

3 de noviembre de 2011

Señores
Comité de Condonación de COLCIENCIAS
Cra 7B Bis No. 132-28
Bogotá D.C
Colombia

Solicitud de Estudio de documentos para
condonación de beca-crédito otorgada por
COLCIENCIAS al
Dr. Leonardo Mariño Ramírez

Actualización : 22 de mayo de 2013



National Institutes of Health
National Library of Medicine
Bethesda, Maryland 20894

3 de noviembre de 2011

Señores
Comité de Condonación de COLCIENCIAS
Cra 7B Bis No. 132-28
Bogotá D.C
Colombia

Solicitud de Estudio de documentos para condonación de beca-crédito otorgada por Colciencias al Dr. Leonardo Mariño Ramírez

Agradezco inmensamente todo el apoyo que he recibido por parte de COLCIENCIAS desde el principio de mi carrera como Investigador. Es mi deseo continuar impulsando el desarrollo de la bioinformática en Colombia a través de actividades de cooperación internacional desde mi posición como Staff Scientist en el National Center for Biotechnology Information (NCBI).

Atentamente me dirijo a Ustedes para solicitarles el estudio de los documentos que pongo a su disposición. En 1997 tuve el honor de ser becario del programa Fulbright-Colciencias-IIE para realizar estudios de doctorado en los Estados Unidos. En 2002 culmine exitosamente mis estudios de doctorado en la Universidad de Texas A&M, regrese a Colombia en 2008 y tengo un programa de investigación activo en bioinformática con varios investigadores Colombianos y pertenezco a dos grupos de investigación clasificados en A (COL0085459) y A1 (COL0078428) respectivamente. Además he participado en la capacitación de un gran numero estudiantes de pregrado y postgrado.

De antemano les agradezco el estudio de mi solicitud de condonación.

Atentamente,

Leonardo Mariño Ramírez, Ph.D.

Staff Scientist, Computational Biology Branch
National Center for Biotechnology Information, NLM, NIH
Building 38A, Room 6S614M
8600 Rockville Pike, MSC 6075
Bethesda, MD 20894-6075
Tel: +1 301-402-3708
E-mail: marino@ncbi.nlm.nih.gov

Leonardo Mariño-Ramírez

Staff Scientist - NCBI / NLM / NIH
Building 38A, Room 6S614M
8600 Rockville Pike
Bethesda, MD 20894
Phone: (301) 402-3708 Fax: (301) 480-2288
E-mail: marino@ncbi.nlm.nih.gov

Education	1997 – 2002	Texas A&M University. College Station, TX. PhD - Biochemistry
	1988 – 1992	Universidad de Los Andes. Bogotá, Colombia BSc - Microbiology
Research Experience	2012 – Present	Computational Biology Branch. National Center for Biotechnology Information. National Library of Medicine. National Institutes of Health. Bethesda, MD Staff Scientist
	2008 – 2012	Computational Biology and Bioinformatics Unit. Biotechnology and Bioindustry Center. Corporación Colombiana de Investigación Agropecuaria (CORPOICA). Bogotá, Colombia. Associate Investigator
	2004 – 2008	Computational Biology Branch. National Center for Biotechnology Information. National Library of Medicine. National Institutes of Health. Bethesda, MD Staff Scientist
	2002 – 2004	Computational Biology Branch. National Center for Biotechnology Information. National Library of Medicine. National Institutes of Health. Bethesda, MD Research Fellow <ul style="list-style-type: none">- Computational analysis of mammalian promoter sequences for the identification of regulatory elements. Supervisor: David Landsman
	1997 – 2002	The Hu lab. Department of Biochemistry and Biophysics. Texas A&M University. College Station, TX. Research Assistant <ul style="list-style-type: none">- Mapping protein oligomerization domains from yeast and <i>E. coli</i> in bacterial cells. Advisor: James C. Hu
	1996	Boyce Thompson Institute for Plant Research. Cornell University. Ithaca, NY. Visiting Scientist <ul style="list-style-type: none">- Genetic transformation of potato (<i>Solanum</i> spp.). Analysis of gene expression during breaking potato

- tuber dormancy.
 Advisor: Charles J. Arntzen
 1995 Institute of Biosciences and Technology. Texas A&M University. Houston, TX.
Visiting Scientist
 - DNA sequence analysis of cucumber mosaic virus (CMV) coat protein gene isolated from Colombian cultivars.
 Advisor: Charles J. Arntzen
 1994 – 1997 National Plant Biotechnology Program. Corporación Colombiana de Investigación Agropecuaria (CORPOICA). Bogotá, Colombia.
Research Assistant
 - Molecular characterization of *Bacillus thuringiensis* isolates.
 - Genetic transformation of banana (*Musa* spp.)
 Advisor: Javier Narvaez-Vasquez
 1992 – 1994 Immunology Institute. San Juan de Dios Hospital. National University of Colombia. Bogotá, Colombia
Research Assistant
 - Molecular characterization of pathogen related genes from *Mycobacterium tuberculosis*.
 Advisor: Manuel E. Patarroyo

Honors and Awards

- DHHS / NIH / NLM – Special Achievement Award (2006)
- GlaxoSmithKline Bioinformatics Prize. Best paper award 13th Annual International Conference on Intelligent Systems for Molecular Biology - ISMB 2005 (2005)
- Encyclopedia Britannica Scholarship (1997)
- Fulbright/Colciencias/IIE pre-doctoral Fellowship (1997)
- Fellowship in Bioinformatics. International Centre for Genetic Engineering and Biotechnology – ICGEB (1996)
- Short-Term Fellowships in Biotechnology. UNESCO (1996)
- Cochran Fellowship Program in Biotechnology. United States Department of Agriculture – USDA (1995)

Memberships in Professional Societies

- American Society for Biochemistry and Molecular Biology (ASBMB)
- International Society for Computational Biology (ISCB)
- Fulbright Alumni Association

- Novoa-Aponte L, León-Torres A, Patiño-Ruiz M, Cuesta-Bernal J, Salazar LM, Landsman D, **Mariño-Ramírez L**, Soto CY. (2012) *In silico* identification and characterization of the ion transport specificity for P-type ATPases in the *Mycobacterium tuberculosis* complex. *BMC Struct Biol.* 12:25.
- Kostka JE, Green SJ, Rishishwar L, Prakash O, Katz LS, **Mariño-Ramírez L**, Jordan IK, Munk C, Ivanova N, Mikhailova N, Watson DB, Brown SD, Palumbo AV, Brooks SC. (2012) Genome sequences for six *Rhodanobacter* strains, isolated from soils and the terrestrial subsurface, with variable denitrification capabilities. *J Bacteriol.* 194:4461-4462.
- Hansen L, **Mariño-Ramírez L**, Landsman D. (2012) Differences in local genomic context of bound and unbound motifs. *Gene.* 506:125-134.
- Spouge JL, **Mariño-Ramírez L**. (2012) The practical evaluation of DNA barcode efficacy. *Methods Mol Biol.* 858:365-377.
- Garzón-Martínez GA, Zhu ZI, Landsman D, Barrero LS, **Mariño-Ramírez L**. (2012) The *Physalis peruviana* leaf transcriptome: assembly, annotation and gene model prediction. *BMC Genomics.* 13:151.
- Heibel SK, Lopez GY, Panglao M, Sodha S, **Mariño-Ramírez L**, Tuchman M, Caldovic L. (2012) Transcriptional regulation of N-acetylglutamate synthase. *PLoS One.* 7:e29527.
- Hansen, L, Kim, NK, **Mariño-Ramírez L**, Landsman D. (2011) Analysis of biological features associated with meiotic recombination hot and cold spots in *Saccharomyces cerevisiae*. *PLoS One.* 6:e29711.
- Huda, A., Tyagi, E., **Mariño-Ramírez, L.**, Bowen NJ, Jjingo D, Jordan IK. (2011) Prediction of transposable element derived enhancers using chromatin modification profiles. *PLoS One.* 6:e27513.
- González-Pérez, M., Murcia, M.I., Landsman, D., I. King Jordan and **Mariño-Ramírez, L.** (2011) The genome sequence of *Mycobacterium colombiense* CECT 3035 type strain. *J. Bacteriol.* 193:5866-5867.
- Simbaqueba, J., Sánchez, P., Sanchez, E., Núñez Zarantes, V. M., Chacon, M. I., Barrero, L. S. and **Mariño-Ramírez, L.** (2011) Development and Characterization of Microsatellite Markers for the Cape Gooseberry *Physalis*

peruviana. PLoS One. 6:e26719.

- **Mariño-Ramírez, L.**, Levine, K. M., Morales, M., Zhang, S., Moreland, R. T., Baxevanis, A. D. and Landsman D. (2011) The Histone Database: an integrated resource for histones and histone fold-containing proteins. Database (Oxford): bar048.
- Jjingo, D., Huda, A., Gundapuneni, M., **Mariño-Ramírez, L.** and I. King Jordan (2011) Effect of the transposable element environment of human genes on gene length and expression. *Genome Biol Evol.* 3:259-271.
- Valderrama-Aguirre, A., Zúñiga-Soto, E., **Mariño-Ramírez, L.**, Moreno, L.Á., Escalante, A.A., Arévalo-Herrera, M. and Herrera S. (2011) Polymorphism of the Pv200L fragment of merozoite surface protein-1 of *Plasmodium vivax* in clinical isolates from the Pacific coast of Colombia. *Am J Trop Med Hyg.* 84(2 Suppl):64-70.
- Huda, A., **Mariño-Ramírez, L.** and Jordan, I. K. (2010) Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mob DNA.* 1:2.
- Hansen, L., **Mariño-Ramírez, L.** and Landsman, D. (2010) Many sequence-specific chromatin modifying protein-binding motifs show strong positional preferences for potential regulatory regions in the *Saccharomyces cerevisiae* genome. *Nucleic Acids Research.* 38:1772-1779.
- Wang, J., Bowen, N. J., **Mariño-Ramírez, L.** and Jordan, I. K. (2009) A c-Myc regulatory subnetwork from human transposable element sequences. *Mol Biosyst.* 5:1831-1839.
- Huda, A., **Mariño-Ramírez, L.**, Landsman, D. and Jordan, I. K. (2009) Repetitive DNA elements, nucleosome binding and human gene expression. *Gene.* 436:12-22.
- Ogurtsov, A. Y., **Mariño-Ramírez, L.**, Johnson, G. R., Landsman, D., Shabalina, S. A. and Spiridonov, N. A. (2008) Expression patterns of protein kinases correlate with gene architecture and evolutionary rates. *PLoS ONE.* 3:e3599.
- Kim, N-K., Tharakaraman, K., **Mariño-Ramírez, L.** and Spouge, J. L. (2008) Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. *BMC Bioinformatics.* 9:262.
- Polavarapu, N., **Mariño-Ramírez, L.**, Landsman, D., McDonald, J. F. and Jordan, I. K. (2008) Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. *BMC Genomics.* 9:226.
- Tharakaraman, K., Bodenreider, O., Landsman, D., Spouge J. L. and **Mariño-Ramírez, L.** (2008) The biological

function of some human transcription factor binding motifs varies with position relative to the transcription start site. *Nucleic Acids Research*. 36:2777-2786.

- Resch, A. M., Carmel, L., **Mariño-Ramírez, L.**, Ogurtsov, A. Y., Shabalina, S. A., Rogozin, I. B. and Koonin, E. V. (2007) Widespread Positive Selection in Synonymous Sites of Mammalian Genes. *Molecular Biology and Evolution*. 24:1821-1831.

- Piriyaongsa, J., **Mariño-Ramírez, L.** and Jordan, I. K. (2007) Origin and evolution of human microRNAs from transposable elements. *Genetics*. 176:1323-1337.

- Riz, I., Akimov, S. S., Eaker, S. S., Baxter, K. K., Lee, H. J., **Mariño-Ramírez, L.**, Landsman, D., Hawley, T. S. and Hawley, R. G. (2007) TLX1/HOX11-induced hematopoietic differentiation blockade. *Oncogene*. 26:4115-4123.

- **Mariño-Ramírez, L.**, Jordan, I. K. and Landsman, D. (2006) Multiple independent evolutionary solutions to core histone gene regulation. *Genome Biology*. 7:R122.

- **Mariño-Ramírez, L.**, Bodenreider, O., Kantz, N. and Jordan, I. K. (2006) Co-evolutionary Rates of Functionally Related Yeast Genes. *Evolutionary Bioinformatics*. 2:295-300.

- Tsaparas, P., **Mariño-Ramírez, L.**, Bodenreider, O., Koonin, E. V. and Jordan, I. K. (2006) Global similarity and local divergence in human and mouse gene co-expression networks. *BMC Evolutionary Biology*. 6:70.

- **Mariño-Ramírez, L.**, Tharakaraman, K., Sheetlin, S. L., Landsman, D. and Spouge, J. L. (2006) Scanning sequences after Gibbs sampling to find multiple occurrences of functional elements. *BMC Bioinformatics*. 7:408.

- **Mariño-Ramírez, L.** and Jordan, I. K. (2006) Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biology Direct*. 1:20.

- **Mariño-Ramírez, L.**, Hsu, B., Baxevanis, A. D. and Landsman, D. (2006) The Histone Database: a comprehensive resource for histones and histone fold-containing proteins. *Proteins*. 62:838-842.

- Eriksson, P. R., Mendiratta, G., McLaughlin, N., Wolfsberg, T., **Mariño-Ramírez, L.**, Pompa, T., Jainerin, M., Landsman, D., Shen, C-H. and Clark, D. J. (2005) Global regulation by the yeast Spt10 protein is mediated through chromatin structure and the histone UAS elements. *Molecular and Cellular Biology*. 25: 9127-9137.

- **Mariño-Ramírez, L.**, Tharakaraman, K., Sheetlin, S., Landsman, D. and Spouge, J. L. (2005) Alignments anchored on genomic landmarks can aid in the identification of

- regulatory elements. *Bioinformatics*. 21 Suppl 1:i440-i448.
- Jordan, I. K., **Mariño-Ramírez, L.** and Koonin, E. V. (2005) Evolutionary significance of gene expression divergence. *Gene*, 345:119-126.
 - **Mariño-Ramírez, L.**, Spouge, J. L., Kanga, G. C. and Landsman, D. (2004) Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res.*, 32:949-958.
 - Jordan, I. K., **Mariño-Ramírez, L.**, Wolf, Y. I. and Koonin, E. V. (2004) Conservation and co-evolution in the scale-free human gene co-expression network. *Molecular Biology and Evolution*. 21:2058-2070.
 - **Mariño-Ramírez, L.**, Minor, J. L., Reading, N. and Hu, J. C. (2004) Identification and mapping of self-assembling protein domains encoded by the *Escherichia coli* K-12 genome using λ repressor fusions. *J. Bacteriol.*, 186:1311-1319.
 - **Mariño-Ramírez, L.** and Hu, J. C. (2002) Isolation and mapping of self-assembling protein domains encoded by the *Saccharomyces cerevisiae* genome. *Yeast*, 19:641-650.
 - Moon, Y. S., Clendennen, S. K., **Mariño-Ramírez, L.** and May, G. D. (1997) Differential gene expression during the break in potato tuber dormancy. *Plant Physiol.*, 114: 1636-1636 Suppl. S.
 - **Mariño-Ramírez, L.** (1997) Clonación del gen de la cápside proteica de una cepa colombiana del virus del mosaico del pepino (CMV) para su expresión en plantas por transformación mediante *Agrobacterium*. *Revista Corpoica* 2, 58-59.
 - Hernández Fernández, J., **Mariño-Ramírez, L.**, Orozco Cárdenas, M. L. y Narváez Vásquez J. (1996) Uso de la reacción en cadena de la polimerasa para la caracterización de aislamientos nativos de *Bacillus thuringiensis*. *Revista Corpoica* 2, 1-9.
 - **Mariño-Ramírez, L.**, Hernández Fernández, J., Orozco Cárdenas, M. L. y Narváez Vásquez J. (1996) Caracterización Molecular de Genes cry de *Bacillus thuringiensis* utilizando PCR Extra-Rápida. *Revista Corpoica* 1, 47-47.
 - Reichel, H., **Mariño-Ramírez, L.**, Kummert, J., Belalcazar, S. y Narváez, J. (1996) Caracterización del gen de la proteína de la cápside de dos aislamientos del virus del mosaico del pepino (CMV), obtenidos de plátano y banano (*Musa* spp.) *Revista Corpoica* 1, 1-5.

Invited Review articles		<p>- Mariño-Ramírez, L., Kann, M. G., Shoemaker, B. A. and Landsman, D. (2005) Histone structure and nucleosome stability. <i>Expert Review of Proteomics</i>. 2:719-729.</p> <p>- Mariño-Ramírez, L., Lewis, K. C., Landsman, D. and Jordan, I. K. (2005) Transposable elements donate lineage-specific regulatory sequences to host genomes. <i>Cytogenetic and Genome Research</i>. 110:333-341.</p>
	Invited Book chapters	<p>- Mariño-Ramírez, L., Tharakaraman, K., Bodenreider, O., Spouge, J. L. and Landsman, D. (2009) Identification of cis-Regulatory Elements in Gene Co-expression Networks Using A-GLAM, in <i>Methods in Molecular Biology: Computational Systems Biology</i>, (McDermott, J.; Samudrala, R.; Bumgarner, R.; Montgomery, K.; Ireton, R. ed.) 541:3-22, Springer, New York, NY.</p> <p>- Mariño-Ramírez, L., Tharakaraman, K., Spouge, J. L. and Landsman, D. (2009) Promoter analysis: gene regulatory motif identification with A-GLAM, in <i>Methods in Molecular Biology: Bioinformatics for DNA Sequence Analysis</i>, (Posada, D. ed.) 537:263-276, Springer, New York, NY.</p> <p>- Jordan, I. K. and Mariño-Ramírez, L. (2007) Evolutionary genomics of gene expression, in <i>Structural Approaches to Sequence Evolution Molecules, Networks, Populations</i> (Bastolla, U.; Porto, M.; Roman, H.E. and Vendruscolo, M. ed.), Springer, New York, NY.</p> <p>- Mariño-Ramírez, L., Campbell, L. and Hu, J. C. (2003) Screening peptide/protein libraries fused to the λ repressor DNA-binding Domain in <i>E. coli</i> cells, in <i>Methods in Molecular Biology: E. coli Gene Expression Protocols</i>, (Vaillancourt, P. ed.) 205:235-250, Humana Press, Totowa, NJ.</p> <p>- Mariño-Ramírez, L. and Hu, J. C. (2002) Using λ repressor fusions to isolate and characterize self-assembling domains, in <i>Protein-Protein Interactions: A Laboratory Manual</i>, (Golemis, E., and Serebriiskii, I. ed.) 375-394, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.</p>
Teaching Experience	2012	- Lecturer: Practical Course "Bioinformatics: Computer Methods in Molecular Biology" - International Centre for Genetic Engineering and Biotechnology (ICGEB). Trieste, Italy
	2008 – Present	- Lecturer: Computational Genomics - School of Biology - Georgia Institute of Technology. Atlanta, Georgia
	2007	- Lecturer: Curso de Bioinformática: Fundamentos para el manejo y uso de datos biológicos. Montería, Colombia

	2006	- Lecturer: Curso Internacional: Manejo de Herramientas Básicas en Bioinformática. Bogotá, Colombia
	2005	- Lecturer: Curso Internacional de Bioinformática: Manejo de las herramientas básicas. Bogotá, Colombia
	2003 – Present	Study group leader: NCBI Perl programming language study group
Academic Service	2011 – Present	Editor: PLoS ONE (Public Library of Science)
	2011 – Present	Steering Committee – NIH Staff Scientists/Staff Clinicians Organization
	2010 – Present	Founding Member: PanAmerican Bioinformatics Institute. (http://panambioinfo.org/)
	2008 – Present	Editor: DATABASE (Oxford Journals)
	2008 – Present	Member of the External Advisory Board for the Professional MS Bioinformatics Program. School of Biology. Georgia Institute of Technology
	2007 – Present	<i>Ad hoc</i> Grant Reviewer: The National Science Foundation (NSF)
	2007 – Present	Editor: GENE (Elsevier)
	2007 – 2008	Steering Committee – NIH Staff Scientists/Staff Clinicians Organization
	2006 – 2009	Program Committee – Intelligent Systems for Molecular Biology (ISMB)
	2005 – Present	- Reviewer for papers in scientific meetings: ISMB, Pacific Symposium on Biocomputing (PSB), VirtualGenomics
	2005 – Present	- <i>Ad hoc</i> Grant Reviewer: The Kentucky Science and Engineering Foundation (KSEF)
	2005 – Present	- Reviewer for scientific journals: Bioinformatics, Nucleic Acids Research (Oxford University Press) Biochemistry (American Chemical Society) Gene, Genomics (Elsevier) Genome Dynamics (Karger) BioMed Central Biomédica (Instituto Nacional de Salud – Colombia)
Research Projects	2002 – Present	Analysis of Gene Regulatory Sequences from Whole Chromosomes and Genomes. Funded by NIH Intramural research project Z01 LM000084.
	2002 – Present	Structural and Functional Analysis of Protein Sequence Families. Funded by the NIH Intramural research project Z01 LM000071.
	2002 – Present	Analysis of Repeated Elements in the Human Genome. Funded by the NIH Intramural research project Z01 LM000092.

2003 – Present	The Analysis of Signal Elements in Promoter Sequences. Funded by the NIH Intramural research project Z01 LM091704.
2001 – 2002	Protein Self-Assembly in Model Microorganisms. Funded by the NIH grant R01GM063652.

**Software
developed**

2005	The Histone Database - http://research.nhgri.nih.gov/histones/
2005	A-GLAM - ftp://ftp.ncbi.nih.gov/pub/spouge/papers/archive/AGLAM/
2002	Doodle Database - http://dimer.tamu.edu/doodle/

Títulos profesionales

Microbiología, pregrado – Universidad de Los Andes - 1992

Bioquímica, doctorado – Texas A&M University – 2002

UNIVERSIDAD DE LOS ANDES

SANTAFE DE BOGOTA, D.C.

REPUBLICA DE COLOMBIA
MINISTERIO DE EDUCACION NACIONAL

El Consejo Directivo y el Rector de la Universidad de los Andes
con las debidas autorizaciones legales y teniendo en cuenta que

Leonardo Mariño Ramírez

C.C. 79'520.882 DE BOGOTA

ha cumplido con los requisitos académicos exigidos por la Universidad, le otorgan con
los derechos, obligaciones y prerrogativas correspondientes, el presente Diploma de

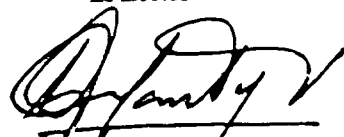
Microbiólogo

ALCALDIA MAYOR DE SANTAFE DE BOGOTA, D.C.

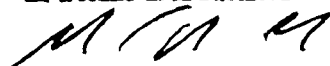
Anotado al Folio No 77-4 del Libro No 92

EL SECRETARIO DE EDUCACION

El Rector



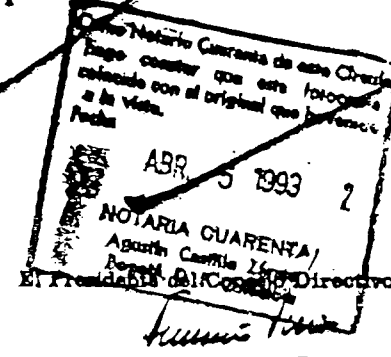
El Decano de la Facultad



REGISTRADO
LIBRO 9 FOLIO 3

Santafé de Bogotá, D.C., a 01 de Octubre de 1992

10217



El Secretario General



Santafé de Bogotá, D.C., 12 de septiembre de 1992

Texas A & M University

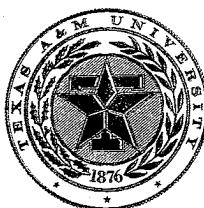
To all to whom these presents may come Greeting
Be it Known that


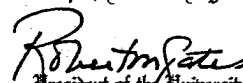
Leonardo Marino-Ramirez

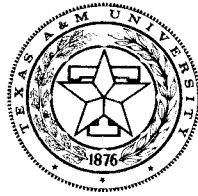
having completed the studies and satisfied the requirements for the Degree of
Doctor of Philosophy

has accordingly been admitted to that Degree with all the honors, rights and
privileges belonging thereto.

Given under the seal of the University at College Station, Texas, on the
twenty-first day of December, A.D., two thousand and two.




Chair, Board of Regents

President of the University



TEXAS A&M UNIVERSITY
Office of the Vice President for Research
Office of Graduate Studies
1113 TAMU • College Station, Texas 77843-1113
(979) 845-3631
FAX (979) 845-1596

22 OCTOBER 2002

TO WHOM IT MAY CONCERN:

This is to certify that the student named below has completed all requirements for the degree indicated.

LEONARDO MARINO-RAMIREZ

NAME

DOCTOR OF PHILOSOPHY

DEGREE

BIOCHEMISTRY

MAJOR

21 DECEMBER 2002

CONFERRAL DATE

A handwritten signature in black ink, appearing to read "John R. Giardino".

John R. Giardino
Dean of Graduate Studies

Certificados de vinculación laboral en Colombia (Regreso al país)

Corporación Colombiana de Investigación Agropecuaria - Corpoica

Centro Colombiano de Genómica y Bioinformática de Ambientes Extremos – Gebix

Fulbright – Colombia

Fundación Instituto de Inmunología de Colombia - FIDIC

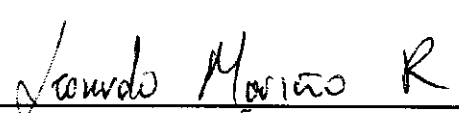
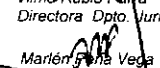
**CONTRATO INDIVIDUAL DE TRABAJO A TÉRMINO INDEFINIDO CON
SALARIO INTEGRAL, CELEBRADO EN LA CORPORACION COLOMBIANA DE
INVESTIGACIÓN AGROPECUARIA, CORPOICA**

Entre los suscritos a saber, **ARTURO ENRIQUE VEGA VARÓN**, identificado con la cédula de ciudadanía No. 6.892.249 de Montería, actuando en su condición de Director Ejecutivo de la Corporación Colombiana de Investigación Agropecuaria CORPOICA, Entidad de participación mixta, con personería jurídica otorgada por la Alcaldía Mayor de Bogotá, según Resolución especial No. 141 del 7 de abril de 1993, quien en adelante se denominará CORPOICA, por una parte y por la otra **LEONARDO MARIÑO RAMIREZ**, mayor de edad, identificado como aparece al pie de su firma, quien en adelante se llamará EL EMPLEADO, se ha celebrado el contrato individual de trabajo que consta de las siguientes Cláusulas: **PRIMERA.** - EL EMPLEADO se obliga a laborar para la Corporación a partir del **01 de septiembre de 2008**, como **Investigador Ph.D Asociado - Laboratorio de Genética Molecular y Bioindustria - CBB**, con cargo al centro de costos **(311-1240-001-1460-00)**, desarrollando las actividades que para dicho cargo señala el manual de descripción de cargos y naturaleza de las dependencias y en la forma que le indique su superior inmediato. **SEGUNDA.** - La sede de EL EMPLEADO será en el **Centro de Investigación Tibaitatá - Mosquera. Parágrafo.** No obstante lo anterior, podrán convenir las partes que el trabajo se preste en lugar distinto del inicialmente contratado, siempre que tales traslados no desmejoren las condiciones laborales o de remuneración del empleado o implique perjuicios para él. Los gastos que se originen con el traslado serán cubiertos por la Corporación de conformidad con el numeral 8o. del artículo 57 del Código Sustantivo del Trabajo. El trabajador se obliga a aceptar los cambios de cargo u oficio o lugar de trabajo que decida la Corporación dentro de su poder subordinante siempre que respete las condiciones laborales del empleado y no se le causen perjuicios, todo ello sin que se afecte el honor, la dignidad, y los derechos mínimos del empleado de conformidad con el artículo 23 del Código Sustantivo del Trabajo, modificado por el artículo 1o. de la Ley 50 de 1990. **TERCERA.** - EL EMPLEADO acepta los Reglamentos de concurso de méritos y los sistemas de evaluación que adopte la Corporación para permanecer en el empleo, así como los respectivos reglamentos de trabajo y seguridad de la Corporación. Las partes están de acuerdo en que todos los reglamentos y sistemas de evaluación de la Corporación forman parte de este contrato. **CUARTA.** - La Corporación pagará a EL EMPLEADO por la prestación de sus servicios la suma de **Seis Millones Quinientos Diez Mil Pesos mensuales (\$6.510.000)** mensuales, como **salario integral**, el cual retribuirá el trabajo ordinario y compensará de antemano el valor de prestaciones, recargos o beneficios tales como el correspondiente al trabajo nocturno extraordinario o al dominical y festivo, el de primas legales, extralegales, cesantías y sus intereses, subsidios y suministros en especie, y en gastos generales, las que se incluyan para los empleados de la

Corporación, excepto las vacaciones. **QUINTA.**-El empleado se obliga a laborar la jornada ordinaria en los términos y dentro de las horas señaladas por la Corporación, pudiendo ésta hacer ajustes o cambios de horario cuando lo estime conveniente. Por el acuerdo expreso o tácito de las partes podrán repartirse las horas de la jornada ordinaria en la forma prevista en el artículo 164 del Código Sustantivo del Trabajo modificado por el artículo 23 de la Ley 50 de 1990, teniendo en cuenta que los tiempos de descanso entre las secciones de la jornada no se computan dentro de la norma según el artículo 167 del Código Sustantivo del Trabajo. **SEXTA.**- Los primeros **sesenta (60) días** del presente contrato se consideran como período de prueba y, por consiguiente, cualquiera de las partes podrá terminar el contrato unilateralmente, en cualquier momento durante dicho período. Vencido este, la duración del contrato será a término **Indefinido**, mientras subsistan las causas que le dieron origen y la materia del trabajo. **SEPTIMA.**- EL EMPLEADO se compromete a respetar todos los derechos de propiedad intelectual y/o industrial sobre los resultados que se obtengan en desarrollo de las actividades derivadas de su vínculo laboral con la Corporación. **Parágrafo Primero:** De conformidad con el artículo 539 del Código de Comercio y demás disposiciones generales y especiales pertinentes, los derechos sobre propiedad intelectual, es decir, las patentes de invención, los descubrimientos, modelos de utilidad y diseños industriales, etc, son de propiedad exclusiva de la Corporación, en consecuencia Corpoica, tendrá el derecho de hacer patentar a su nombre o a nombre de terceros, los inventos o mejoras, respetándose el derecho del empleado a ser mencionado como inventor en la patente, si así lo desea, sin que Corpoica quede obligada al pago de compensación alguna. **Parágrafo Segundo:** El EMPLEADO se compromete a suscribir el Contrato de Cesión de Derechos Patrimoniales de Autor, como también todos aquellos documentos, medios y poderes necesarios, que la Corporación requiera para adelantar las gestiones ante la Superintendencia de Industria y Comercio o la Entidad que haga sus veces, a efectos de obtener los derechos de propiedad intelectual o beneficios comerciales, derivados del desarrollo de las actividades como trabajador de Corpoica. **Parágrafo Tercero:** El EMPLEADO se obliga a no divulgar comunicaciones o publicaciones orales o escritas, que tiendan a impedir o entorpecer la petición de patentes o registros, salvo autorización expresa y escrita otorgada por la Corporación. **OCTAVA.**- EL EMPLEADO se compromete a mantener confidencialidad sobre toda la información científica, técnica, financiera, comercial o de cualquier otra índole, que le sea suministrada o que conozca directa o indirectamente en desarrollo de las actividades como trabajador de la Corporación Colombiana de Investigación Agropecuaria, Corpoica. **NOVENA.**- Son justas causas para dar por terminado este contrato las enumeradas en el Código Sustantivo del Trabajo y además por parte de la Corporación las siguientes que para tal efecto se califican como graves: a) La violación por parte del empleado de cualquiera de sus obligaciones legales contractuales o reglamentarias y especialmente las contenidas en el numeral 5 del artículo 69 del Reglamento Interno de Trabajo de la Corporación aprobado por el Ministerio de Trabajo y

Seguridad Social. b) No obtener en las evaluaciones periódicas de sus actividades la calificación exigida en los reglamentos. c) La ejecución por parte del empleado de labores remuneradas al servicio de terceros sin autorización escrita de la Corporación. d) La revelación o utilización de secretos y datos reservados de la Corporación a los que tenga acceso con ocasión del desempeño de sus labores. e) Presentarse embriagado o ingerir bebidas embriagantes en el sitio de trabajo aún por primera vez. **DECIMA.-** El presente contrato reemplaza en su integridad y deja sin efecto cualquier otro contrato verbal o escrito, celebrado entre las partes con anterioridad.

Para constancia se firma en Bogotá, a los **veintidós (22) días del mes de febrero de 2008.**

EL EMPLEADOR,**EL EMPLEADO,**
ARTURO ENRIQUE VEGA VARÓN
Corpoica
LEONARDO MARIÑO RAMÍREZ
C.C. No. 79'520.882Vo. Bo. 
Carlos Fernando Ortiz Gómez
Secretario GeneralVo. Bo. 
Paula Buitrago Galiano
Subdirectora Financiera NacionalVo. Bo. 
Vilma Rubio Parra
Directora Dpto. JurídicoVo. Bo. 
Jairo Ortiz Castellanos
Director Dpto. de Gestión HumanaElabora: 
Marlén Gaitana Vega
Auxiliar Activo II - Dpto. Gestión Humana



**CONTRATO DE PRESTACION DE SERVICIOS No. 031-2009
CELEBRADO ENTRE LA UNION TEMPORAL CENTRO COLOMBIANO DE
GENÓMICA Y BIOINFORMÁTICA DE AMBIENTES EXTREMOS – GeBiX Y
LEONARDO MARIÑO RAMIREZ**

Entre los suscritos **MARIA MERCEDES ZAMBRANO EDER**, mayor de edad, identificada con cédula de ciudadanía No. 31.874.035 expedida en Cali, quien obra en nombre y representación de la **UNION TEMPORAL CENTRO COLOMBIANO DE GENÓMICA Y BIOINFORMÁTICA DE AMBIENTES EXTREMOS – GeBiX**, y quien para los efectos del presente contrato se denominará **GeBiX** y **LEONARDO MARIÑO RAMIREZ**, identificado con cédula de ciudadanía No. 79.520.882 expedida en Bogotá, quien obra en su propio nombre y quien en adelante se llamará **EL CONTRATISTA**, han acordado celebrar el presente Contrato de Prestación de Servicio que se registrá por las siguientes cláusulas:

PRIMERA.- OBJETO: **EL CONTRATISTA** se compromete a prestar sus servicios profesionales como Profesional en Bioinformática para asesorar a GEBIX en el montaje de su plataforma Bioinformática, en el marco del proyecto “Conformación de una plataforma en metagenómica y bioinformática para la caracterización y el aprovechamiento de recursos genéticos de ambientes extremos” ejecutado por GeBiX.

SEGUNDA.- OBLIGACIONES DEL CONTRATANTE:

1. Suministrar oportunamente la información y documentación requerida por **EL CONTRATISTA** a fin de dar cumplimiento a lo estipulado en este contrato.
2. De requerirse, permitir y hacer los esfuerzos necesarios para que el contratista se reúna e interactúe oportunamente con los directores de los grupos de investigación de GeBiX.
3. Desembolsar lo acordado dentro de este Contrato de Prestación de Servicios, en la forma y fechas señaladas, previa presentación de la respectiva cuenta de cobro.
4. Suministrar al **CONTRATISTA**, en caso de ser necesario, pasajes aéreos o terrestres dentro del país, estadía y demás gastos de viaje, que requiera para el cumplimiento de las obligaciones adquiridas en el presente Contrato.

TERCERA.- OBLIGACIONES DEL CONTRATISTA:

1. Brindar asesoría general a los grupos de GEBIX trabajando en bioinformática.
2. Ayudar en la definición de los flujos de trabajo para análisis de diversidad y para análisis de secuencias metagenómicas.
3. Asesorar en el análisis de los datos metagenómicos generados.
4. Asistir a las reuniones programadas por el grupo de bioinformática de Gebix y, de ser necesario, a reuniones generales de GEBIX.
5. Estar en disposición de trasladarse temporalmente fuera de su domicilio contractual, en caso de ser necesario.
6. Afiliarse y aportar a una Entidad Promotora de Salud y a un Fondo de Pensiones como independiente durante la duración del presente contrato.
7. Presentar las respectivas cuentas de cobro.



CUARTA.- EXCLUSIONES: El contratista no asumirá ninguna responsabilidad por eventuales infracciones o violaciones a derechos de propiedad intelectual en que pudieren incurrir las personas jurídicas y naturales que conforman la Unión Temporal GeBiX. El contrato no implica una asesoría jurídica general, es decir en las diversas áreas del derecho; lo anterior, sin perjuicio de que el contratista colabore emitiendo opiniones, y conceptos en las materias consultadas al Grupo de Investigación PLEBIO relacionadas con el objeto de la Unión Temporal GeBiX.

SEXTA.- DURACIÓN: El presente Contrato de Prestación de Servicios tendrá una duración de ocho (8) meses a partir de la suscripción del contrato.

SÉPTIMA.- VALOR DEL CONTRATO Y FORMA DE PAGO: El valor del presente Contrato de Prestación de Servicios es la suma de OCHO MILLONES NOVECIENTOS OCHENTA Y CINCO MIL SEISCIENTOS PESOS (\$8,985.600,00) MONEDA CORRIENTE, que se cancelarán en ocho (8) desembolsos mensuales, cada uno de UN MILLON CIENTO VEINTITRES MIL DOSCIENTOS PESOS (\$1.123.200,00) MONEDA CORRIENTE, sujetos a la entrega de un informe de actividades. El último desembolso estará sujeto a la entrega de un informe final de las actividades realizadas, junto con los soportes generados de las obligaciones del contrato. **PARAGRAFO 1.-** Cada uno de los desembolsos requiere para su cancelación de la presentación de la cuenta de cobro acompañada de la certificación sobre el cumplimiento de las obligaciones a cargo del **CONTRATISTA**, otorgada por el supervisor del contrato, así como del recibo vigente de cotización al Sistema General de Salud y Pensiones. **PARAGRAFO 2.- GeBiX** pagará la suma acordada en este contrato dentro de los treinta (30) días siguientes a la presentación de la cuenta de cuenta de cobro por parte del **CONTRATISTA**. **PARAGRAFO 3.- GeBiX** deja claro que de conformidad con la legislación vigente al respecto, las personas naturales que presten sus servicios a las entidades del sector privado bajo la modalidad de contrato de prestación de servicios, están obligados a afiliarse y aportar al Sistema General de Salud y Pensiones. **PARÁGRAFO 4.-** En el evento de ser necesario el desplazamiento de **EL CONTRATISTA** a una ciudad diferente, por razón o en ocasión de las actividades contratadas, **GeBiX** asumirá los gastos de viaje, viáticos y pasajes para realizar las actividades mencionadas anteriormente. Para el desembolso de los gastos de desplazamiento, **EL CONTRATISTA** deberá presentar la solicitud con visto bueno del ordenador del gasto, dentro de los cinco (5) días hábiles siguientes al regreso, **EL CONTRATISTA**, deberá legalizar los gastos de viaje anexando los soportes correspondientes.

OCTAVA.- SUSPENSIÓN TEMPORAL: De común acuerdo entre las partes, se podrá suspender la ejecución del contrato, mediante la suscripción de acta, sin que para efectos del plazo extintivo del contrato se compute el tiempo de suspensión, siempre y cuando esta suspensión no exceda el 30% del tiempo pactado para la asesoría. En caso de suspenderse el contrato por cualquier causa, **EL CONTRATISTA** tendrá derecho a la remuneración que se hubiere causado hasta la fecha de la suspensión, la que se determinará de conformidad con el servicio que hubiere realizado hasta esa fecha.

NOVENA.- TERMINACIÓN ANTICIPADA O PRORROGA: El presente contrato podrá darse por terminado por mutuo acuerdo entre las partes, expresado por escrito, o en forma unilateral por cualquiera de ellas, manifestando la parte que desea hacerlo a la otra, su intención de darlo por terminado, con una antelación no



inferior a treinta (30) días, por incumplimiento de cualquiera de las obligaciones derivadas del contrato, o en uno cualquiera de los casos siguientes: **1.** Suspensión total o parcial del proyecto sin justa causa imputable a la otra parte; **2.** Incumplimiento de una o cualquiera de las obligaciones que asumen los contratantes; **3.** Por acuerdo mutuo entre las partes; **4.** Por no afiliación o desafiliación de las entidades de seguridad social como independiente por parte de **EL CONTRATISTA**; **5.** Por realización de programa de estudios en el exterior.

DÉCIMA.- EL CONTRATISTA será responsable de los daños y perjuicios que se causen a **GeBiX** o a terceros, con motivo de la ejecución de los trabajos contratados, cuando resulten de: **1.** Incumplimiento a los términos y condiciones establecidos en el presente Contrato; **2.** Inobservancia a las recomendaciones que **GeBiX** le haya dado por escrito; **3.** Actos con dolo, mala fe o negligencia; **4.** La pérdida de material, o detrimento grave en los equipos que le hayan sido dados para el correcto desarrollo de la labor contratada; **5.** En general por actos u omisiones graves imputables a **EL CONTRATISTA** o al personal que este llegare a emplear.

DÉCIMA PRIMERA.- INDEPENDENCIA DEL CONTRATISTA: EL CONTRATISTA actuará por su propia cuenta, con absoluta autonomía y no estará sometida a subordinación laboral alguna con **GeBiX**. Sus derechos se limitarán, de acuerdo con la naturaleza del presente Contrato de Prestación de Servicios a exigir el pago de lo estipulado por la prestación de sus servicios. Queda claramente entendido que no existirá relación laboral alguna entre **GeBiX** y **EL CONTRATISTA**, o el personal que ésta utilice en la ejecución del objeto del presente Contrato.

DÉCIMA SEGUNDA.- SUPERVISIÓN: El presente Contrato de Prestación de Servicios estará bajo la Supervisión del Dr. HOWARD ARMANDO JUNCA DIAZ PhD o quien haga sus veces. Dicha supervisión comprende, entre otras, el recibo y la verificación de los servicios contratados.

DÉCIMA TERCERA.- CESIÓN DEL CONTRATO: EL CONTRATISTA no podrá ceder parcial ni totalmente la ejecución del presente Contrato a un tercero salvo previa autorización expresa y escrita de **GeBiX**.

DECIMA CUARTA.- PROPIEDAD INTELECTUAL: Los derechos patrimoniales de autor sobre los documentos, ya sean éstos impresos o archivos magnéticos o electrónicos, generados como resultado de la ejecución del objeto del presente contrato de prestación de servicios, pertenecerán a **GeBiX**, quién podrá, en consecuencia, publicarlos, reproducirlos, cederlos y en general disponer de ellos a cualquier título. Para estos efectos el contratista se compromete a cumplir las formalidades requeridas para la transferencia de derechos patrimoniales conforme a la legislación vigente.

DÉCIMA QUINTA.- CONFIDENCIALIDAD: EL CONTRATISTA se comprometa a guardar absoluta reserva acerca de toda la información relativa a los procedimientos y procesos técnicos o científicos que en desarrollo del objeto o de las funciones adelante **GeBiX**. Esta obligación de confidencialidad estará a cargo del **CONTRATISTA** durante el término de duración del contrato y un (1) año más. En consecuencia, **EL CONTRATISTA** se obliga a no utilizar ni divulgar para fines distintos a los previstos en este contrato los resultados de su trabajo conseguidos en la



ejecución del mismo, como tampoco la información que conozca con ocasión del contrato, sin la previa autorización expresa y escrita que para cada caso reciba de **GeBiX** y de **COLCIENCIAS**, pues se considera que todos los documentos empleados e información que se produzca directa y exclusivamente en desarrollo del presente contrato pertenecerán, en cuanto a derechos patrimoniales y de explotación económica se refiere, a las instituciones que conforman la Unión Temporal en proporción a su participación real efectiva en la realización de los trabajos que resulten patentables. La confidencialidad a que se refiere esta cláusula se mantendrá hasta que la información adquiera el carácter de pública.

Sin embargo el contratista podrá utilizar los resultados de su trabajo como insumos para análisis o estudios de caso en proyectos de investigación y publicación académicos. En todo caso deberá dar los reconocimientos institucionales correspondientes e informar a GeBiX, con copia de las publicaciones.

DÉCIMA SEXTA.- SOLUCIÓN DE CONFLICTOS: Las partes convienen que en el evento en que surja alguna diferencia entre las mismas, por razón o en ocasión del presente Contrato, en que deba acudirse a la justicia ordinaria, previamente se acudirá al Centro de Conciliación de la Cámara de Comercio de Bogotá.

DÉCIMA SÉPTIMA.- DOMICILIO CONTRACTUAL: Para todos los efectos legales a que hubiera lugar, las partes acuerdan como domicilio contractual la ciudad de Bogotá, y las notificaciones serán recibidas por las partes en las siguientes direcciones: Por **GeBiX**, en la carrera 5 No. 66A-34 y por **EL CONTRATISTA** en la Carrera 28A No. 50-77 Apto. 301, en la ciudad de Bogotá.

DÉCIMA OCTAVA.- REQUISITOS DE PERFECCIONAMIENTO: El presente Contrato se perfeccionará con la suscripción del mismo por las partes, y la presentación del Registro Único Tributario – RUT por parte del **CONTRATISTA**.

De conformidad con lo anterior, las partes suscriben el presente documento, en dos copias del mismo valor y tenor, en la ciudad de Bogotá D.C., a los veinte (20) día del mes de abril de 2009.

Por **GeBiX**

EL CONTRATISTA

MARIA MERCEDES ZAMBRANO EDER
Representante Legal

Leonardo Mariño R.
LEONARDO MARIÑO RAMIREZ
C.C. No. 79.520.882 de Bogotá

Bogotá D.C., 13 de septiembre de 2010

No. 00521

Doctor
LUIS RICARDO FERNANDEZ RESTREPO
Coordinador
Oficina de Asuntos Consulares
Ministerio de Relaciones Exteriores
Ciudad

Asunto: OAJ No. 66042

De manera atenta confirmamos que, el pasado 15 de enero de 2007 respondimos el oficio de referencia OAJ No. 66042, en el cual informábamos que el señor **LEONARDO MARIÑO RAMIREZ**, identificado con c.c. 79.520.882, fue beneficiario del Programa Fulbright Colciencias – IIE, año 1997. En dicha comunicación, también manifestábamos que el señor Mariño Ramirez debía cumplir con el requisito estipulado por el Programa Fulbright de permanecer en Colombia por un período mínimo de dos años, después de la fecha de terminación de su programa en Estados Unidos.

Posterior a la mencionada comunicación, el 04 de septiembre de 2010, recibimos los soportes a través de los cuales se certifica que el señor Mariño Ramirez regresó a Colombia después de finalizar su beca y trabajó con la Corporación Colombiana de Investigación Agropecuaria - Corpoica de forma continua desde septiembre de 2008 hasta la actualidad a través de una contrato indefinido. Por lo tanto, informamos al Ministerio de Relaciones Exteriores, que el señor **LEONARDO MARIÑO RAMIREZ** ya cumplió con el requisito estipulado y anteriormente descrito, y por lo tanto a esta fecha, Fulbright Colombia No Objeta la estadia de este ciudadano en los Estados Unidos.

Atentamente,



ANN C. MASON
Directora Ejecutiva

Elaboró LMDCC

c.c. Embajada de Estados Unidos
c.c. Leonardo Mariño Ramirez



FUNDACION INSTITUTO DE INMUNOLOGIA DE COLOMBIA FIDIC

NIT 830084143-6

DAFF-3300-263-0099

Abril 15 de 2013

LA DIRECTORA ADMINISTRATIVA Y FINANCIERA DE LA FUNDACIÓN INSTITUTO DE INMUNOLOGIA DE COLOMBIA –FIDIC–

CERTIFICA

Que el(a) BIOINFORMÁTICO **LEONARDO MARIÑO RAMIREZ**, identificado(a) con CC No. 79.520.882 Expedida en Bogotá, estuvo vinculado(a) a esta institución mediante la modalidad de Contrato de Prestación de Servicios Profesionales Independientes, como **Consultor** del Centro, bajo las siguientes condiciones:

Contrato No. 2010-0098

Contrato de Prestación de Servicios Profesionales Independientes

Objeto: Prestar los servicios de consultoría y capacitación en el área de Bioinformática para el montaje y expansión su plataforma Bioinformática, en el marco del proyecto "Validación de una prueba serológica para identificación de anticuerpos anti-VPH e infección persistente por el virus de papiloma humano en mujeres de escasos recursos y víctimas del conflicto armado"

Grupo Funcional: Biología Molecular e Inmunología

Fecha de Inicio: Octubre 1 de 2010

Fecha de Terminación: Noviembre 30 de 2010

Obligaciones del consultor:

- Brindar asesoría general a los científicos que conforman los grupos funcionales que requieran de sus servicios trabajando en bioinformática.
- Ayudar en la definición de los flujos de trabajo para análisis de secuencias presentes en el genoma de Plasmodium vivax y P. alci parum de interés inmunológico.
- Asesorar en el análisis de los datos generados.
- Asistir a las reuniones programadas por los grupos de bioinformática, según disponibilidad.
- Dedicación de 10 horas semanales.

Contrato No. 2011-0153

Contrato de Prestación de Servicios Profesionales Independientes

Objeto: Prestar los servicios de consultoría y capacitación en el área de Bioinformática para el montaje y expansión su plataforma Bioinformática, en el marco del proyecto "Plan de Fortalecimiento Institucional 2011-2013"

Grupo Funcional: Biología Molecular e Inmunología

Fecha de Inicio: Marzo 21 de 2011

Fecha de Terminación: Octubre 20 de 2011



FUNDACION INSTITUTO DE INMUNOLOGIA DE COLOMBIA FIDIC

NIT 830084143-6

**Obligaciones del
consultor:**

- Brindar asesoría general a los científicos que conforman los grupos funcionales que requieran de sus servicios trabajando en bioinformática.
- Ayudar en la definición de los flujos de trabajo para análisis de secuencias presentes en el genoma de Plasmodium vivax y P. alciparum de interés inmunológico.
- Asesorar en el análisis de los datos generados.
- Asistir a las reuniones programadas por los grupos de bioinformática, según disponibilidad.
- Dedicación de 10 horas semanales.

Contrato No. 2012-0012

Contrato de Prestación de Servicios Profesionales Independientes

Objeto: Prestar los servicios de consultoría y capacitación en el área de Bioinformática para el montaje y expansión su plataforma Bioinformática, en el marco del proyecto "Plan de Fortalecimiento Institucional 2011-2013"

Grupo Funcional: Biología Molecular e Inmunología

Fecha de Inicio: Enero 16 de 2012

Fecha de Terminación: Noviembre 15 de 2012

**Obligaciones del
consultor:**

- Brindar asesoría general a los científicos que conforman los grupos funcionales que requieran de sus servicios trabajando en bioinformática.
- Ayudar en la definición de los flujos de trabajo para análisis de secuencias presentes en el genoma de Plasmodium vivax y P. alciparum de interés inmunológico.
- Asesorar en el análisis de los datos generados.
- Asistir a las reuniones programadas por los grupos de bioinformática, según disponibilidad.
- Dedicación de 10 horas semanales.

Esta Certificado se expide a solicitud del (a) interesado(a) para tramites personales.

MARIA O. JIMÉNEZ P.

Directora Administrativa y Financiera

C.C. Hoja de vida del contratista
Archivo de Certificaciones contratistas FIDIC

Grupo Funcional Biología Molecular e Inmunología, MAPG

Publicaciones - 2007-2012

Genome Sequence of the *Mycobacterium colombiense* Type Strain, CECT 3035

Mónica González-Pérez,^{1,2,3} Martha I. Murcia,^{1,2} David Landsman,³
I. King Jordan,^{2,4} and Leonardo Mariño-Ramírez^{1,2,3*}

Departamento de Microbiología, Facultad de Medicina, Universidad Nacional de Colombia, Bogotá, Colombia¹; PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia²; Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894³; and School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332⁴

Received 2 August 2011/Accepted 4 August 2011

We report the first whole-genome sequence of the *Mycobacterium colombiense* type strain, CECT 3035, which was initially isolated from Colombian HIV-positive patients and causes respiratory and disseminated infections. Preliminary comparative analyses indicate that the *M. colombiense* lineage has experienced a substantial genome expansion, possibly contributing to its distinct pathogenic capacity.

The genus *Mycobacterium* comprises nearly 150 species (2, 3), including a number of human pathogens that pose major challenges to public health. *Mycobacterium colombiense* is a slow-growing, urease-positive, nontuberculous mycobacterium (NTM) that belongs to the *Mycobacterium avium* complex (MAC). *M. colombiense* was originally isolated from HIV-positive individuals in Bogotá, Colombia, and the patient isolates were determined to represent a distinct species by virtue of sequence comparisons with closely related *Mycobacterium* species (7). Since the discovery of this new species in 2006, *M. colombiense* has been confirmed to cause respiratory disease and disseminated infection in immunocompromised HIV patients, as well as lymphadenopathy in immunocompetent children (1, 8). Nevertheless, very little is currently known about the molecular mechanisms that underlie *M. colombiense* infection and pathogenesis. We have characterized the complete genome sequence of *M. colombiense* in an effort to better understand its virulence mechanisms.

The *M. colombiense* genome was sequenced by a whole-genome shotgun strategy using Roche 454 GS-FLX titanium pyrosequencing technology. A total of 720,174 sequence reads were generated, with an average read length of 375 bp, yielding more than 270 Mb of total sequence. This represents 45× coverage for the estimated 5.6-Mb genome size. A *de novo* assembly of the 454 single-end data was created using the Newbler assembler (Roche), version 2.6, resulting in 27 large contigs with an N₅₀ of 436 kb. Genome annotation was performed using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP), which produces functional annotation using the NCBI nonredundant protein and protein cluster databases with functional domain assignments for each protein by RPS-BLAST (5) against the NCBI Conserved Domain Database (6). The *M. colombiense*

genome was predicted to encode 5,230 coding sequences (CDS).

M. colombiense was previously shown to be most closely related to *M. avium*, based on 16S rRNA sequence analysis along with DNA-DNA hybridization experiments (7). Here, we show that *M. colombiense* is most closely related to *M. avium* subsp. *paratuberculosis* (4) and confirm these results via sequence comparisons of *M. colombiense* contigs against the NCBI microbial sequence database. Despite the close relationship between these two species, reference-based assembly of the *M. colombiense* genome using *M. avium* subsp. *paratuberculosis* produced a highly fragmented assembly, with markedly lower quality than seen for the *de novo* assembly (1,914 large contigs with an N₅₀ of 1,253), indicating that numerous genome rearrangements have occurred since the two species diverged. Furthermore, our characterization of the *M. colombiense* genome shows it to be substantially larger (5.6 Mb) than the genome of *M. avium* (4.8 Mb) and to encode many more genes (5,230 versus 4,400). Sequence alignments between the two species revealed that these differences could be attributed to large genomic insertions specific to the *M. colombiense* lineage. We hypothesize that a genome expansion may have allowed for the elaboration of novel pathways that contribute to the virulence of this emerging opportunistic pathogen. Additional genomic and functional analyses are needed to interrogate this hypothesis.

Nucleotide sequence accession number. The *M. colombiense* Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number AFVW000000000.

This work was supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839 to I.K.J.) and the NCBI Scientific Visitors Program (ORISE to M.G.-P.). The research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI.

We thank the Spanish Type Culture Collection (CECT) for providing strains.

* Corresponding author. Mailing address: Computational Biology Branch, Building 38A, Room 6S614M, 8600 Rockville Pike, MSC 6075, Bethesda, MD 20894-6075. Phone: (301) 402-3708. Fax: (301) 480-2288. E-mail: marino@ncbi.nlm.nih.gov.

REFERENCES

1. **Esparcia, O., F. Navarro, M. Quer, and P. Coll.** 2008. Lymphadenopathy caused by *Mycobacterium colombiense*. *J. Clin. Microbiol.* **46**:1885–1887.
2. **Euzéby, J. P.** 2011, posting date. List of Prokaryotic names with Standing in Nomenclature, genus *Mycobacterium*. J. P. Euzéby, Société de Bactériologie Systématique et Vétérinaire (SBSV), France. <http://www.bacterio.cict.fr/m/mycobacterium.html>.
3. **Euzéby, J. P.** 1997. List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. *Int. J. Syst. Bacteriol.* **47**:590–592.
4. **Li, L., et al.** 2005. The complete genome sequence of *Mycobacterium avium* subspecies paratuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **102**:12344–12349.
5. **Marchler-Bauer, A., and S. H. Bryant.** 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* **32**:W327–W331.
6. **Marchler-Bauer, A., et al.** 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**:D225–D229.
7. **Murcia, M. I., E. Tortoli, M. C. Menendez, E. Palenque, and M. J. Garcia.** 2006. *Mycobacterium colombiense* sp. nov., a novel member of the *Mycobacterium avium* complex and description of MAC-X as a new ITS genetic variant. *Int. J. Syst. Evol. Microbiol.* **56**:2049–2054.
8. **Vuorenmaa, K., I. Ben Salah, V. Barlogis, H. Chambost, and M. Drancourt.** 2009. *Mycobacterium colombiense* and pseudotuberculous lymphadenopathy. *Emerg. Infect. Dis.* **15**:619–620.

Effect of the Transposable Element Environment of Human Genes on Gene Length and Expression

Daudi Jjingo¹, Ahsan Huda¹, Madhumati Gundapuneni^{1,2}, Leonardo Mariño-Ramírez^{3,4}, and I. King Jordan^{*,1,4}

¹School of Biology, Georgia Institute of Technology

²Institute for Systems Biology, Seattle, Washington

³National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

⁴PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia

*Corresponding author: E-mail: king.jordan@biology.gatech.edu.

Accepted: 22 February 2011

Abstract

Independent lines of investigation have documented effects of both transposable elements (TEs) and gene length (GL) on gene expression. However, TE gene fractions are highly correlated with GL, suggesting that they cannot be considered independently. We evaluated the TE environment of human genes and GL jointly in an attempt to tease apart their relative effects. TE gene fractions and GL were compared with the overall level of gene expression and the breadth of expression across tissues. GL is strongly correlated with overall expression level but weakly correlated with the breadth of expression, confirming the selection hypothesis that attributes the compactness of highly expressed genes to selection for economy of transcription. However, TE gene fractions overall, and for the L1 family in particular, show stronger anticorrelations with expression level than GL, indicating that GL may not be the most important target of selection for transcriptional economy. These results suggest a specific mechanism, removal of TEs, by which highly expressed genes are selectively tuned for efficiency. MIR elements are the only family of TEs with gene fractions that show a positive correlation with tissue-specific expression, suggesting that they may provide regulatory sequences that help to control human gene expression. Consistent with this notion, MIR fractions are relatively enriched close to transcription start sites and associated with coexpression in specific sets of related tissues. Our results confirm the overall relevance of the TE environment to gene expression and point to distinct mechanisms by which different TE families may contribute to gene regulation.

Key words: gene expression, gene regulation, selection hypothesis, genomic design hypothesis, L1, MIR.

Introduction

The relationship between gene architecture and gene expression has been and remains a subject of continuing interest for genome analysis. In a pioneering study, Castillo-Davis et al. (2002) observed that, for human and worm genes, intron length was negatively correlated with the level of expression. In other words, shorter genes were found to be expressed at higher levels and longer genes at lower levels. To explain this trend, the authors formulated the “selection hypothesis” (Castillo-Davis et al. 2002). This hypothesis posits that highly expressed genes are shorter due to selective forces that operate in favor of minimizing the energy and time expended during tran-

scription. Subsequently, the relationship between gene length (GL) and expression level was confirmed by a number of studies, providing support for the selection hypothesis (Eisenberg and Levanon 2003; Urrutia and Hurst 2003; Comerón 2004; Chen et al. 2005; Seoighe et al. 2005; Li et al. 2007).

In 2004, Vinogradov (2004) also observed that compact genes were more highly expressed, but he offered a different explanation for this trend. Vinogradov proposed the “genomic design” hypothesis, which postulates that the shorter length of highly expressed genes is better explained by the fact that these genes also tend to be broadly expressed across numerous tissues and thus have simpler regulation, and require fewer regulatory sequence elements, than

genes expressed in a more narrow tissue-specific fashion. In other words, the relative paucity of regulatory elements in broadly expressed genes explains their shorter average length. The genomic design hypothesis rests on the notion that the apparent correlation between GL and the level of expression actually reflects a relationship between GL and the breadth of expression, that is, the number of tissues in which a gene is expressed.

The selection hypothesis and the genomic design hypothesis make distinct testable predictions regarding the relationship between GL and gene expression. The selection hypothesis predicts the strongest correlation between GL and the overall expression level, whereas the genomic design hypothesis predicts the strongest correlation between GL and the breadth of expression. A recent study used these predictions to evaluate the competing hypotheses and found that the selection hypothesis serves as the best explanation for the relationship between GL and expression (Carmel and Koonin 2009).

While the aforementioned studies were ongoing, there was an independent line of research investigating the relationship between gene architecture and gene expression from a different perspective. In eukaryotic genomes, and particularly for mammalian genomes, gene architecture is substantially influenced by the presence of transposable element (TE)-derived sequences. TE-derived sequences are extremely abundant in mammalian genomes; at least 45% of the human genome is made up of TE sequences (Lander et al. 2001; Venter et al. 2001). In addition, TE sequences are nonrandomly distributed across genomes. In the human genome, Alu (SINE) elements are enriched in GC- and gene-rich regions, whereas L1 (LINE) elements are enriched in low-GC and gene-poor regions (Smit 1999; Lander et al. 2001). Finally, individual genes can vary tremendously with respect to the amount and identity of TE sequences that they harbor.

Over the last several years, a series of studies have called attention to a relationship between the TE environment in and around genes and the level and breadth of gene expression. In 2003, the human genome sequence was used together with expression data to construct a human transcriptome map (Versteeg et al. 2003). This map identified colocated clusters of highly expressed genes with specific genomic characteristics. These clusters were gene dense, had high GC content, were enriched for SINEs, Alu elements in particular, and had low LINE densities. The same study found clusters of weakly expressed genes with low SINE and high LINE densities. Shortly thereafter, Han et al. (2004) confirmed that the most highly expressed human genes were depleted for L1 elements and demonstrated a mechanism that could partially explain this pattern. They showed that L1 elements can disrupt transcriptional elongation based on the presence of strong polyA signals in their sequences.

Kim et al. made an important contribution to this body of work by distinguishing between TE effects on the level of expression and the breadth of expression (Kim et al. 2004). They measured overall expression level as the peak expression (PE) over all tissues and breadth of expression (BE) as the number of tissues in which a gene is expressed over some basal threshold. Their work revealed that Alu element gene densities are more highly correlated with BE, whereas L1 densities are most negatively correlated with PE. These results suggested that different families of TEs may have specific effects on different aspects of gene expression. Consistent with these results, Eller et al. showed that highly and broadly expressed housekeeping genes can be distinguished by their TE content, being primarily enriched for Alus and depleted for L1s (Eller et al. 2007). In addition to the level and breadth of expression, the TE environment of mammalian genes has also been related to expression in cancer tissues (Lerat and Semon 2007) and the evolutionary divergence of gene expression (Pereira et al. 2009).

As of yet, no one has attempted to consider these two areas of investigation together: 1) the relationship between GL and expression and 2) the relationship between TE environment and gene expression. In this study, we attempt to disentangle the effects of GL and TE environment on gene expression and to evaluate the relative influences of each on expression. Having considered their effects separately, we then more thoroughly evaluate the connections between gene architecture and the selection versus genomic design hypotheses.

Materials and Methods

Defining Gene Loci

To accommodate alternative splice variants of human genes and compute TE fractions for specific loci, we define genes here as distinct transcriptional units (TUs)—genomic regions encompassing all overlapping transcripts from the start of the 5'-most exon to the end of the 3'-most exon (supplementary fig. S1A, Supplementary Material online). To that end, we downloaded RefSeq annotations for the March 2006 build of the human genome reference sequence (National Center for Biotechnology Information [NCBI] build 36.1; University of California–Santa Cruz [UCSC] hg18) from the UCSC Genome Browser (Karolchik et al. 2004; Rhead et al. 2010). A total of 32,128 RefSeq transcripts were merged into 19,123 TUs that represent distinct gene loci.

Determining Genic and Intergenic TE Fractions

To determine the fractions of human genes (TUs) that are made up of TE sequences, human TEs were broken down into six of the major human TE classes or families according to the Repbase classification system (Jurka et al. 2005; Kohany et al. 2006)—Alu, MIR, L1, L2, DNA and LTR (long terminal repeat). RepeatMasker ([Downloaded from \[gbe.oxfordjournals.org\]\(http://gbe.oxfordjournals.org/\) at NIH Library on October 14, 2011](http://</p>
</div>
<div data-bbox=)

www.repeatmasker.org) annotations of the genomic coordinates of these TEs were used to map them onto their colocated genes. For each TE type, its fraction in a gene was computed as the number of base pairs occupied by a TE as a fraction of all base pairs in the gene. For each human gene, its intergenic region was taken as the union of the regions upstream of the transcription start site (TSS) and downstream of the termination site to the genomic midpoint between the adjacent upstream and downstream genes. TE intergenic fractions were then calculated in the same way as for TE genic fractions based on these genomic coordinates.

Gene Expression Data

To measure gene expression in different tissues, we used the Gene Expression Atlas from the Genomics Institute of the Novartis Research Foundation, which consists of Affymetrix microarray gene expression values for 44,776 probe sets across 79 human tissues (Su et al. 2004). Affymetrix probe sets were mapped onto their corresponding TUs based on their genomic location coordinates. As suggested previously (Stalteri and Harrison 2007), probes that mapped to more than one TU were discarded, and for TUs with more than one mapped probe, the average expression level per tissue was used. This resulted into a final data set of 15,658 TUs to which expression data could be assigned. Expression data are represented as signal intensity units based on the Affymetrix MAS4 processing and normalization algorithm suite.

Measurement of GL and Gene Expression Parameters

For each TU, the GL was calculated by simply subtracting its start coordinate along the chromosome from the end coordinate and then subjecting the difference to a log₂ transformation. The microarray expression data described above were used to calculate three measurements of gene expression: peak expression (PE), breadth of expression (BE) and tissue-specificity (TS). To obtain PE, the signal intensity value from the tissue where the TU is most highly expressed was selected for each TU and subjected to a log₂ transformation to accommodate the vast disparity (range = 197,652.4 signal intensity units) in the peak levels of expression between TUs. For each TU, the BE was calculated as the number of tissues in which the expression of the TU exceeded a threshold of 350 expression signal intensity units (Jordan et al. 2005). For each TU, a TS index was computed as described (Yanai et al. 2005). The value of TS varies between 0 and 1 and reflects the number of tissues where the TU is overly expressed relative to its expression in other tissues. The TS index is calculated as follows:

$$TS = \frac{\sum_{i=1}^N (1 - x_i)}{N - 1},$$

where N is the number of tissues and x_i represents a TU's signal intensity value in each tissue i divided by the maximum signal intensity value of the TU across all tissues.

Comparative Analysis of GL, TE Gene Fractions, and Gene Expression Parameters

The relative effects of GL and the TE gene environment on gene expression were evaluated using pairwise and multiple linear regression analyses where GL and the TE fractions were the independent variables and the gene expression parameters PE, BE, and TS were the dependent variables. For these analyses, parameter values were ranked and binned in order to smooth the signal and reduce the background noise. For each parameter, the 15,658 TUs were ranked and divided into 100 bins of approximately equal size (~157 TUs per bin). Parameter values were averaged for each bin and the averages were used to populate ordered vectors of values ($n = 100$). Vectors that represent independent and dependent variables were then compared using pairwise regression or combined into a multiple regression model. All data were treated using the same ranking and binning procedure so that the relative effects of the independent variables on the dependent variables could be comparatively evaluated.

Gene Expression Clustering Analysis

TS patterns for the top 10% MIR-rich genes were analyzed using hierarchical clustering based on pairwise Euclidean distances between vectors of tissue-specific gene expression levels over 79 tissues. This analysis was conducted using the program Genesis (Sturn et al. 2002) with signal intensity values median normalized across tissues.

Statistical Analyses Used

For the pairwise regression analyses, independent and dependent variable vectors were compared using pairwise Pearson correlation (r values in *figs. 1–5*; individual coefficient of determination R^2 values in *tables 1–5*), and the significance of the correlations (P values in *figs. 1–5* and *tables 1–5*) was determined using the Student's t -distribution. Partial correlation analyses were used to control for the effects of correlated pairs of independent variables (*tables 1, 2 and 4*). Multiple regression analyses were conducted to determine the combined coefficient of determination for all TE fractions (R^2 values in *table 3*) and the partial correlation values (r values in *table 3*). Significance values for the multiple coefficients of determination ("all TE" P values in *table 3*) were determined using the F distribution. Significance values for the partial correlations (P values in *tables 1–4*) were determined using the Student's t -distribution.

Results and Discussion

TE Environment of Human Genes

Gene and TE annotations from the reference sequence of the human genome (NCBI build 36.1; UCSC hg18) were analyzed together to characterize the TE environment of human genes. A total of 19,123 TUs, which reconcile alternative splice variants and represent discrete gene loci, were derived from RefSeq annotations as described in the Materials and Methods (see also [supplementary fig. S1A](#), Supplementary Material online). The fraction of each human gene locus derived from TE sequences was determined using RepeatMasker annotations. Six of the most abundant classes (families) of TEs were considered in this analysis—Alu, MIR, L1, L2, DNA and LTR. The frequencies of other classes of TEs were found to be too low to substantially affect the overall TE environment of human genes.

Human genes show an average TE fraction of 34% and a standard deviation (SD) of 18% ([fig. 1A](#)). Human TE gene fractions show a broad distribution that is fairly bell shaped with the exception of a sharp peak of genes that are devoid of TEs (0% TE fraction in [fig. 1A](#)). The presence of these TE-free genes is consistent with the removal of genic TEs by purifying selection ([Simons et al. 2006](#)). The TE gene fractions observed for individual TE families are consistent with previous results ([Medstrand et al. 2002](#)) in which Alu elements were found to be the most abundant family of TEs in human genes, whereas LTR elements are found in the lowest frequency within human genes ([supplementary fig. S1B](#), Supplementary Material online). The length distributions of TEs in genes ([supplementary table S2](#), Supplementary Material online) reveal that they are mostly short (<400 bp) as would be expected in transcribed regions where long TEs are less tolerated owing to their higher propensity to be deleterious.

Overall, intergenic regions show higher TE fractions (average = 46%; [fig. 1A](#)) and also have a more normal distribution with lower variation than seen for genic regions (SD = 14%; [fig. 1A](#)). For individual human genes, genic and intergenic TE fractions are highly positively correlated ($r = 0.95$, $P = 6.3 \times 10^{-53}$; [fig. 1B](#)), consistent with the notion that the local genomic environment strongly influences TE gene fractions ([Smit 1999](#); [Lander et al. 2001](#)).

TE Fractions are Related to GL

As noted in the introduction, the relationship between GL and expression has been investigated separately from the relationship between the TE environment of genes and their expression. However, GL and gene TE fractions may be related if genes increase in length due, at least in part, to an accumulation of TE-derived sequences. If genes increase in length due to the acquisition of TEs, then we expect to see a positive correlation between gene TE fractions and GL. On

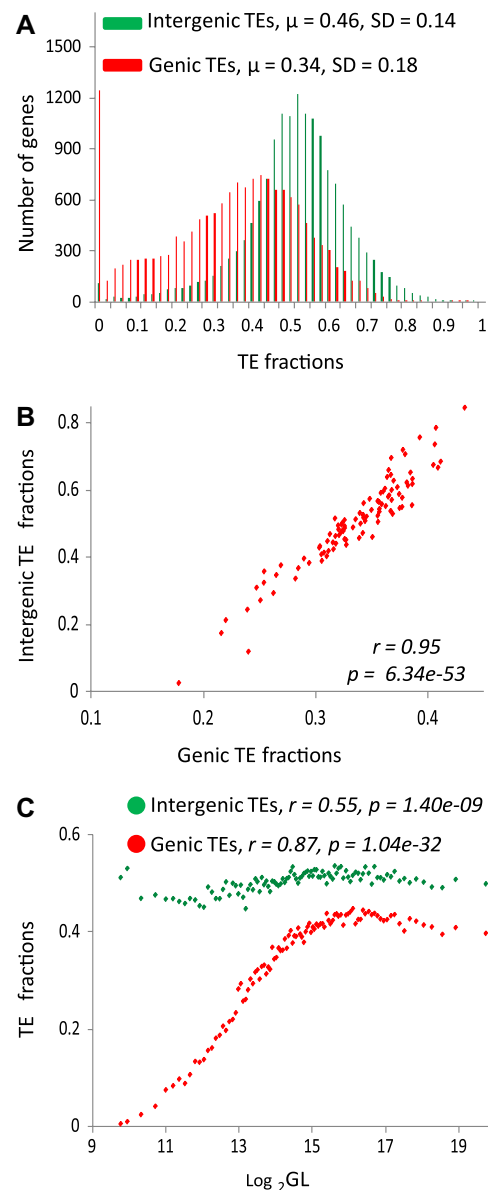


Fig. 1.—TE fractions in and around human genes. (A) Distributions of intergenic (green) and genic (red) TE fractions. (B) Relationship between intergenic TE fractions and the corresponding genic TE fractions. (C) Relationship between intergenic TE fractions and GL (green) and relationship between genic TE fractions and GL (red). Pearson correlation coefficient values (r) along with their significance values (P) are shown for all pairwise regressions.

the other hand, if GL increases via mechanisms that do not involve TEs, there should be no correlation between gene TE fractions and GL. To distinguish between these two possibilities, we compared the TE fractions of human genes with their length (as described in Materials and Methods).

When all human TEs are considered together, there is a strong and significantly positive correlation between gene TE fractions and GL ($r = 0.87$, $P = 1.0 \times 10^{-32}$; [fig. 1C](#)).

Table 1

Relationship between the Local TE Environment and GL

	TE Fractions	<i>r</i>	<i>P</i> Value
GL	Genic TE ^a	0.87	1.04E-32
	Intergenic TE ^a	0.55	1.40E-09
	Genic TE Intergenic TE ^b	0.82	6.80E-45
	Intergenic TE Genic TE ^c	−0.18	7.02E-02

^a TE fractions within genes (genic) and between genes (intergenic) are correlated with GL.

^b Partial correlation between genic TE fractions and GL controlling for intergenic TE fractions.

^c Partial correlation between intergenic TE fractions and GL controlling for genic TE fractions.

Although only 0.55% of the average GL for the bin with the 1 % shortest genes is constituted by TEs, the percentage progressively increases to 39.73% for the bin with the top 1 % longest genes, a >72-fold increase in the average fractions of genes occupied by TEs. However, the positive relationship between gene TE fractions and GL is not strictly monotonic. Specifically, in 77% of all genes, the percentage of GL constituted by TEs progressively increases from 0.55% in genes of about 850 bp to 44.79% for genes spanning about 70.9 kb (>81-fold increase in gene TE fraction; [fig. 1C](#)). For the remaining genes beyond this length (23% of all genes), the percentage of GL constituted by TEs levels off and remains more or less constant with increasing length.

As noted in the previous section, TE genic and intergenic fractions are highly correlated ([fig. 1B](#)). These data are consistent with previous studies showing that TE fractions and family distributions differ among genomic compartments and thus may depend on regional factors such as GC content and recombination rate ([Medstrand et al. 2002](#); [Versteeg et al. 2003](#)). Therefore, it is possible that the relationship between genic TE fractions and GL simply reflects such regional genomic features. To test for this possibility, we compared intergenic TE fractions with GL. Intergenic TE fractions are significantly positively correlated with GL ($r = 0.55$, $P = 1.4 \times 10^{-9}$); however, the correlation is substantially weaker than seen for genic TE fractions and the slope of the relationship is far more flat ([fig. 1C](#)). Furthermore, partial correlation analysis shows that TE genic fractions remain positively correlated with GL when intergenic TE fractions are controlled for, whereas the positive correlation between intergenic TE fractions and GL disappears when genic TE fractions are controlled for ([table 1](#)). In other words, the relationship between TE gene fractions and GL does appear to have some gene-specific, as opposed to genomic regional, component.

To evaluate the correlation between TE genic fractions and GL more closely, we focused on individual TE families and found that Alus dominate the leveling off in gene TE fractions seen for the longest genes. Alus are the most abundant TE sequence within gene boundaries ([supplementary fig. S1B](#), Supplementary Material online), and Alus also

show a unique TE fraction distribution with GL. The fraction of Alus within genes rises sharply and peaks for midsize genes (~23.3 kb) followed by an almost equally precipitous decline in frequency, yielding a bell-shaped distribution ([fig. 2A](#) and [supplementary fig. S3A](#), Supplementary Material online). However, the distribution of TE gene fractions for all other TE families analyzed tends to be generally linear in relation to GL ([fig. 2B](#); [supplementary fig. S3B–F](#), Supplementary Material online), increasing from an average percentage of 0.34% in the shortest genes to 32.83% in the longest genes (a >96-fold increase in the fractions of genes occupied by TEs).

It is not immediately apparent while Alu fractions, unique among all classes of TEs considered here, decline for the longest genes. One possibility is that Alus are known to be prevalent in GC-rich regions, whereas larger genes (introns) tend to have lower GC content ([fig. 2C](#)). Thus, it may be that the decline in Alu content for longer genes is based on regional genomic biases in GC content. If this is the case, then genes with low GC content should also have low Alu fractions and vice versa. We found that genes with low GC content do in fact have lower Alu content as expected ([fig. 2D](#)). However, the relationship between genic Alu fractions and GC content is not monotonic; Alu fractions peak for genes in the middle of the GC content range and decrease for both low- and high-GC content genes. We performed partial correlation in an attempt to further tease apart the relationship between Alu gene fractions and GC content as they relate to GL. GC content is much more strongly correlated with GL than Alu fractions are ([fig. 2A and C](#)). If the relationship of Alu genic fractions with GL mainly reflects regional changes in GC content, then the correlation of Alu fractions with GL should decrease when GC content is controlled for. However, when GC content is controlled for with partial correlation, the positive correlation between Alu gene fractions and GL actually increases ([table 2](#)). Similarly, when Alu gene fractions are controlled for, the correlation between GC content and GL becomes more negative. These data suggest that both Alu gene fractions and GC content are independently related, to some extent, with GL in the human genome.

Overall, the positive correlations between TE gene fractions and GL indicate that longer genes have disproportionately more TEs relative to other sequence elements. Considering all TE families together, TEs make up only 0.55% of the shortest genes and yet account for ~40% of the increase in GL when assessed in the longest genes. For three-fourth of all genes, the contribution of TEs to increases in GL is >45%. These results underscore the contributions of TEs to the length differences among human genes and suggest that the influences of TE environment and GL on gene expression cannot be adequately considered separately.

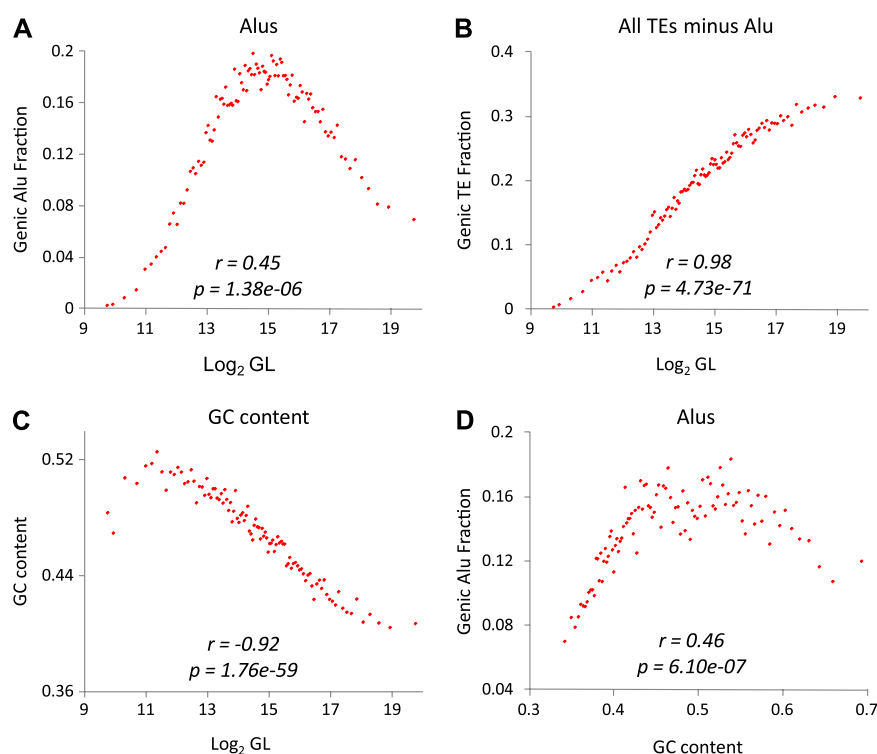


FIG. 2.—Relationships between the Alu fractions of human genes, GL, and GC content. (A) Relationships between Alu gene fractions and GL. (B) Relationship between TE gene fractions for all TEs except Alu and GL. (C) Relationship between GC content and GL. (D) Relationships between Alu gene fraction and GC content. Pearson correlation coefficient values (r) along with their significance values (P) are shown for all pairwise regressions.

TE Gene Environment and the Selection Hypothesis

In order to relate the TE environment of human genes and GL to gene expression, three expression parameters for human genes were measured using microarray data over 79 tissues as described in the Materials and Methods: 1) peak expression (PE), 2) breadth of expression (BE) and 3) TS. PE is the maximum expression level observed for a gene over all 79 tissues and is taken to represent the overall gene expression level; BE is the number of tissues in which a gene can be considered to be expressed, and TS is a measure of tissue specificity described previously (Yanai et al. 2005). PE and BE were measured here because they can be used to distinguish between the selection versus genomic design hypotheses. The selection hypothesis predicts a stronger positive correlation of PE

with GL, whereas the genomic design hypothesis predicts a stronger correlation of BE with GL. However, BE has been criticized as an overly simplistic measure that may not distinguish genes that are expressed in the same sets of tissues albeit at very different relative levels. For this reason, we also use a measure of TS that explicitly reflects the number of tissues where a gene is overly expressed relative to its expression in other tissues (see Materials and Methods). Genes overly expressed in a few tissues (i.e., tissue-specific genes) have high TS indices, whereas more broadly and evenly expressed genes have low values of TS.

Regression analysis was used to individually compare values of these expression parameters with TE gene fractions for all six families and GL (figs. 3–5), and the effects of TE gene fractions and GL were also considered jointly using multiple regression (table 3). Consistent with previous results (Eisenberg and Levanon 2003; Carmel and Koonin 2009), GL can be seen to have a much stronger association with PE than BE. Whereas 48% of the variability in PE is attributable to GL, only about 4% of the variability in BE is attributable to GL (table 3). Furthermore, it can be seen that the nonmonotonic shape of the relationship between GL and PE (fig. 3H) is similar to what has been reported previously (Carmel and Koonin 2009) and also closely resembles the shape of the Alu gene fraction versus PE

Table 2

Effect of GC Content on the Relationship between Alu Genic Fractions and GL

Feature ^a		r	P Value	Control ^b		r	P Value
GL	Alu	0.45	1.32E-06	Alu GC	0.58	1.69E-12	
	GC	−0.92	5.93E-42	GC Alu	−0.94	2.99E-152	

^a Alu genic fractions and genic GC content values are correlated with GL.

^b Partial correlation analyses control for effect of GC content on Alu fractions (Alu | GC) and Alu fractions on GC content (GC | Alu), respectively.

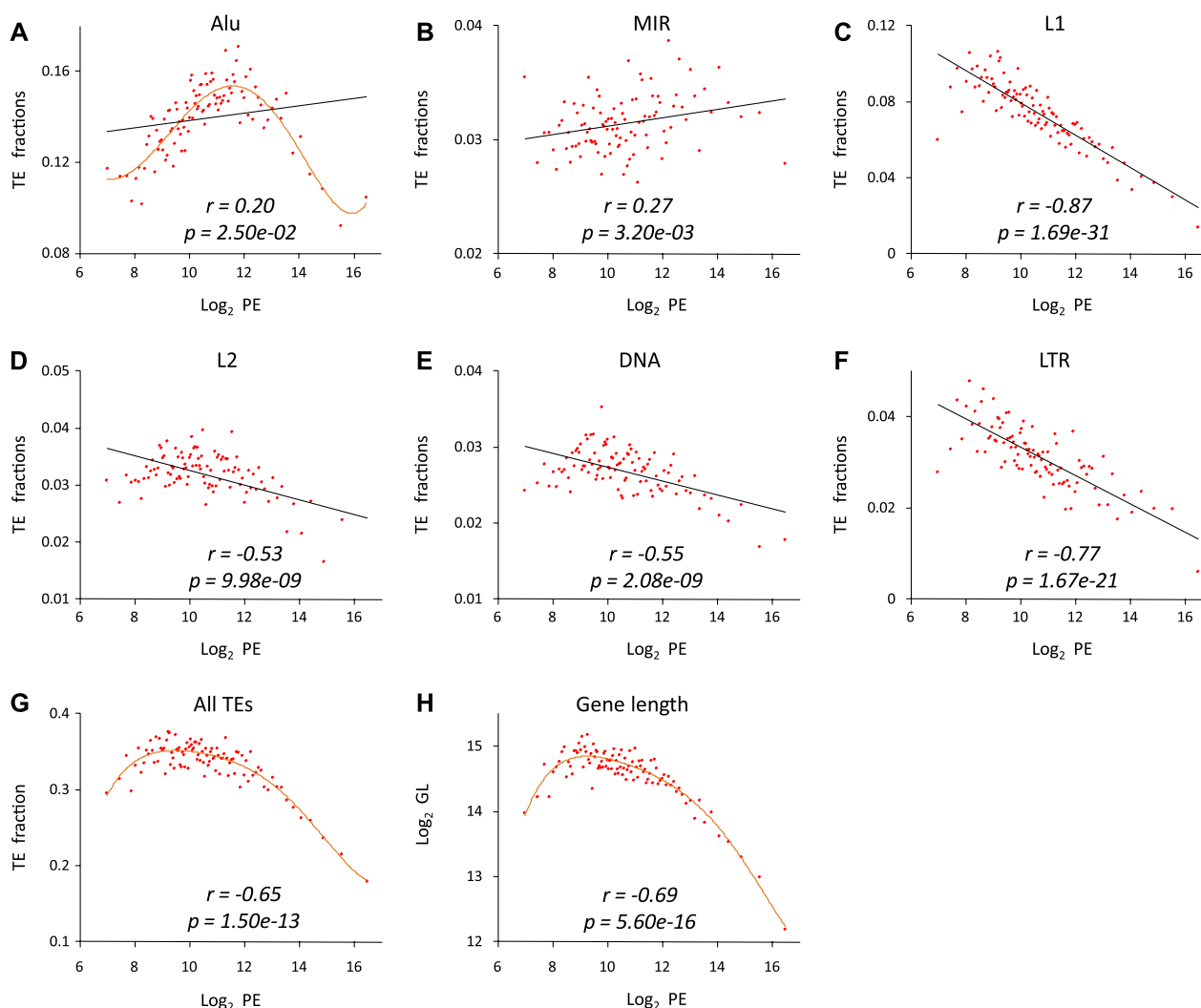


FIG. 3.—TE fractions, GL, and the peak expression (PE). Relationships between the TE gene fractions for (A) Alu, (B) MIR, (C) L1, (D) L2, (E) DNA, (F) LTR, and (G) all TEs and the PE of human genes. (H) Relationship between GL and PE. Pearson correlation coefficient values (r) along with their significance values (P) are shown for all pairwise regressions.

distribution (fig. 3A). The strongest individual TE family correlation with PE is the negative correlation seen for L1 fraction versus PE (fig. 3C). L1 also has the largest negative partial correlation value with PE in the multiple regression analysis as well as the largest coefficient of determination (table 3). When all TEs are analyzed together, 78% of the variability in PE can be attributed to variability in TE gene fractions, whereas only 48% is attributable to variability in GL (table 3).

Although these data do lend support to the selection hypothesis, they also indicate that TE-derived sequences within genes are more highly correlated with their expression level than the overall GL. Thus, the selective mechanism for streamlining highly expressed genes may be related more to the elimination, or shortening, of TE sequences per se rather than the overall shortening of genes.

TE Gene Environment and the Genomic Design Hypothesis

The relationship between GL and BE seen here is generally weak; GL has one of the lower individual correlations with BE (fig. 3G), and variability in GL only contributes 9% of the variability seen in BE (table 1). In addition, the results show that although all the longest genes are narrowly expressed, there are about as many compact narrowly expressed genes as there are compact broadly expressed genes (fig. 4H). Even more surprising is the fact that the partial correlation value for GL versus BE is positive, albeit marginally (table 3), and not negative as can be expected if more narrowly expressed genes are in fact longer.

To interrogate the genomic design hypothesis more closely, we used TS as an alternate measure for the tissue specificity of expression. The genomic design hypothesis

Table 3

The Relationship between TE Fractions, GL, and Gene Expression

Expression Parameter	TE and GL	Coefficient of Determination		Partial Correlation	
		R^2 ^a	P Value	r^b	P Value
PE	All TEs	0.78	<2.2E-16	-0.13	2.1E-01
	L1	0.75	<2.2E-16	-0.86	2.6E-63
	LTR	0.60	<2.2E-16	-0.20	4.5E-02
	GL	0.48	1.1E-15	-0.13	2.2E-01
	DNA	0.29	4.2E-09	-0.01	9.4E-01
	L2	0.27	2.0E-08	-0.25	1.4E-02
	MIR	0.06	6.3E-03	0.25	1.1E-02
	Alu	0.03	5.0E-02	0.32	1.1E-03
BE	All TEs	0.76	<2.2E-16	-0.10	3.1E-01
	Alu	0.59	<2.2E-16	0.52	3.0E-09
	LTR	0.57	<2.2E-16	-0.37	1.0E-04
	L1	0.47	2.8E-15	-0.52	2.4E-09
	MIR	0.12	2.2E-04	-0.28	3.6E-03
	GL	0.04	3.2E-02	0.15	1.5E-01
	L2	0.02	7.4E-02	0.08	4.4E-01
	DNA	0.01	1.3E-01	0.14	1.7E-01
TS	All TEs	0.66	<2.2E-16	-0.32	8.8E-04
	L1	0.63	<2.2E-16	-0.67	9.5E-19
	GL	0.53	<2.2E-16	-0.05	6.3E-01
	L2	0.30	3.0E-09	-0.21	3.3E-02
	Alu	0.29	5.0E-09	-0.13	2.2E-01
	LTR	0.28	9.4E-09	-0.24	1.8E-02
	MIR	0.27	2.1E-08	0.31	1.6E-03
	DNA	0.24	1.8E-07	-0.04	7.3E-01

^a R^2 (the coefficient of determination) is the fraction of variability in each expression parameter that can be attributed to the variability in each sequence feature (individual TE families, GL, or all TEs combined).

^b r is the partial correlation of each feature with the expression parameters, taking into account the presence of the other elements. For each expression parameter, the TEs and GL are ranked by their predictive value for the parameter.

posits that increasing GL is based on the requirement for additional regulatory sequences in genes that are expressed more narrowly. Thus, in the case of TS, a positive correlation is expected between GL and TS; in other words, longer genes are expected to be more tissue specific. For the pairwise regression analysis, there is actually a strongly negative correlation between GL and TS (fig. 5H). This negative trend holds when the TE fractions are controlled for in the partial correlation, and GL also has a high coefficient of determination for TS (table 3). It should be noted that the negative correlation between GL and TS may be related to the analytical formulation used to compute TS (see Materials and Methods) because genes with high expression levels in one or a few tissues (i.e., high PE) will often, but not always, have high TS as well. Nevertheless, when taken together, the data for both GL versus BE and GL versus TS seem to argue against the genomic design hypothesis as originally conceived.

With respect to the TEs, there are strongly positive (Alu; fig. 4A) and negative (L1; fig. 4C) correlations between TE gene fractions and BE, and 76% of the variability in BE can be attributed to variability in all TE gene fractions (table 3).

Overall, TE gene fractions also have the highest coefficient of determination for TS. Consistent with what was previously shown for PE, these data suggest that the combinatorial impact of TEs in human genes is more important than the overall GL with respect to the number of tissues in which a gene is expressed and the tissue specificity of genes.

L1 Elements and Gene Expression Levels

As described previously, the data analyzed here provide support for the selection hypothesis because GL is more strongly (negatively) correlated with PE than BE. However, the strongest negative correlation with PE in the pairwise regression analysis is seen for L1 gene fractions (fig. 3C). L1 also has the highest negative partial correlation with PE in the multiple regression analysis and the highest coefficient of determination (table 3); 75% of the variability in PE is attributable to L1 gene fractions compared with the 48% explained by GL. Thus, L1 gene fractions are more predictive of PE than GL, indicating that variation in the gene fractions of L1s is associated with a higher change in gene expression than variation in GL.

It is also possible that regional genomic features, such as GC content, contribute to the apparent effect of L1 gene content on PE. It is known that L1 elements are enriched in GC-poor regions (Smit 1999; Lander et al. 2001), whereas GC content is strongly positively correlated with PE and BE (Vinogradov 2005). Thus, one may expect to see the kind of negative correlations between L1 and PE/BE seen here based solely on regional biases in GC content. We performed partial correlation to separate the effects of L1 gene fractions and GC content on both PE and BE. When we control for GC content, the partial correlation of L1 fractions with PE remains highly significant (table 4). Conversely, when we control for L1 fractions, the partial correlation of GC with PE is rendered insignificant (table 4). Both L1 fractions and GC content show similar levels of relatedness with BE and partial correlation analysis does not remove either effect (table 4). Thus, the relationship between L1 gene fractions and PE/BE cannot be explained solely by the genomic distribution of L1s among different GC content regions.

L1 elements are an abundant and recently active family of LINES that make up 17% of the human genome sequence (Lander et al. 2001; Venter et al. 2001). Experimental studies have demonstrated that the presence of L1 sequences within genes can lower transcriptional activity (Han et al. 2004; Ustyugova et al. 2006). The effect of the presence of L1s on PE observed here may be attributed to the fact that the disruptive activity of L1s on transcription inhibits gene expression more than an overall increase in GL does. However, this finding is not entirely inconsistent with the selection hypothesis, rather it suggests a specific mechanism, namely the elimination of L1 sequences, for selectively

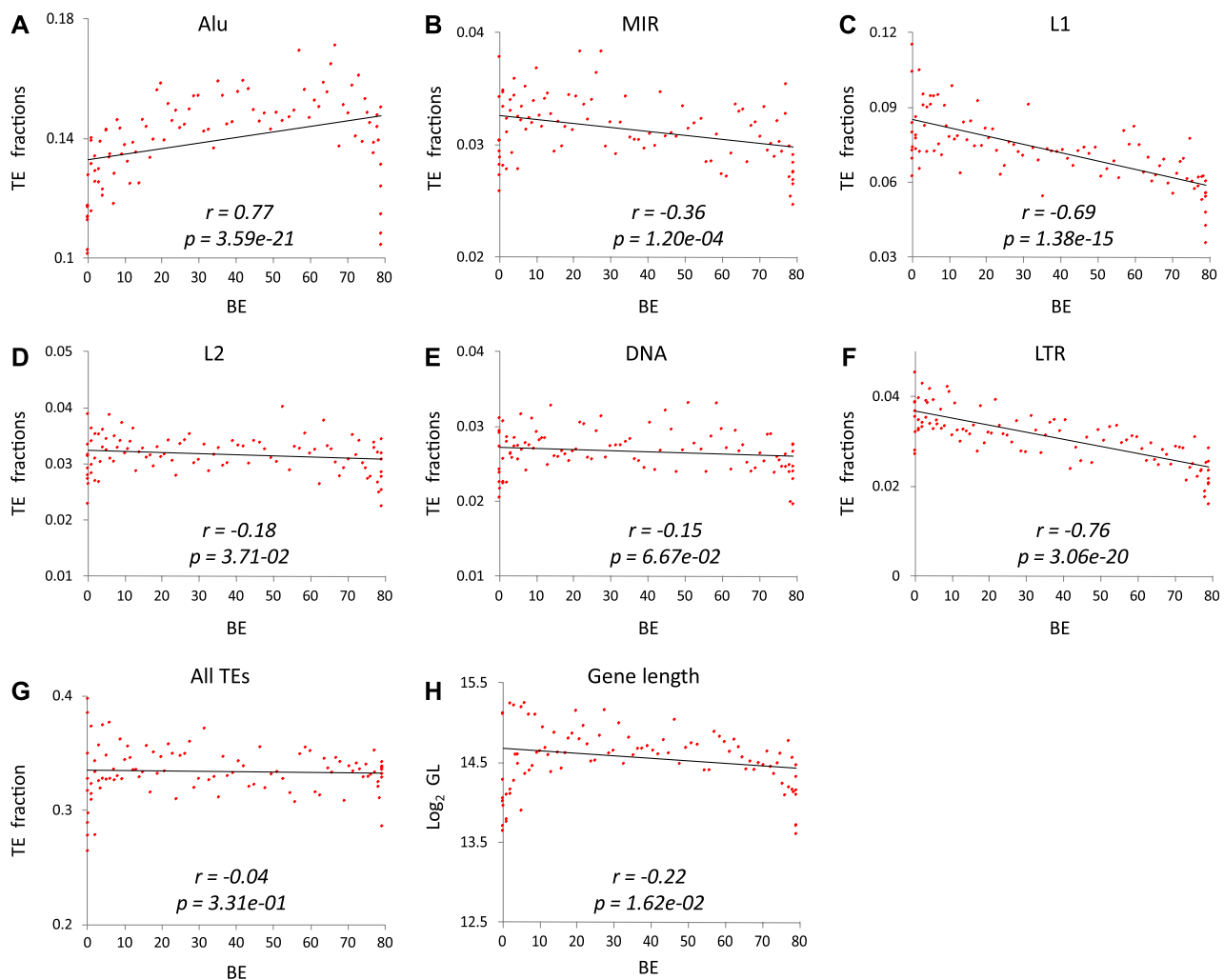


FIG. 4.—TE fractions, GL, and the breadth of expression (BE). Relationships between the TE gene fractions for (A) Alu, (B) MIR, (C) L1, (D) L2, (E) DNA, (F) LTR, and (G) all TEs and the BE of human genes. (H) Relationship between GL and BE. Pearson correlation coefficient values (r) along with their significance values (P) are shown for all pairwise regressions.

tuning highly expressed genes that would also result in an overall decrease in their length.

MIR Elements and Tissue-Specific Gene Expression

The genomic design hypothesis posits a requirement for additional regulatory sequence elements that facilitate TS, which in turn leads to an increase in GL. However, data reported here show that the presence of such regulatory elements does not necessarily result in an overall increase in GL as predicted by the genome design hypothesis (fig. 5H). In light of this realization, we sought to evaluate whether any specific TE sequence elements might be related to the regulatory complexity entailed by tissue-specific genes. Of all the TE families evaluated, MIRs are the only elements that show the expected trends for the genome design hypothesis for both BE and TS. The fraction of MIRs in human genes is negatively correlated with BE (fig. 4B) and positively

correlated with TS (fig. 5B) as expected. In fact, MIRs are the only TEs positively correlated with TS, and the increase in the MIR gene fraction is not linear with increasing TS. At the high range of TS (>0.7 ; 58% of all genes), the positive correlation of MIR gene fractions to TS is even stronger ($r = 0.78$, $P = 3.7 \times 10^{-18}$).

These results are interesting in light of what is already known about MIRs. MIR elements (mammalian-wide interspersed repeats) are an ancient family of transfer RNA-derived SINEs (Jurka et al. 1995; Smit and Riggs 1995), and they have previously been implicated as having regulatory significance in a number of studies. Initially, human MIR sequences were shown to be highly conserved over time suggesting that they may encode some unknown regulatory function (Silva et al. 2003). Subsequently, MIR-derived sequences have been shown to donate transcription factor-binding sites (Polavarapu et al. 2008; Wang et al. 2009),

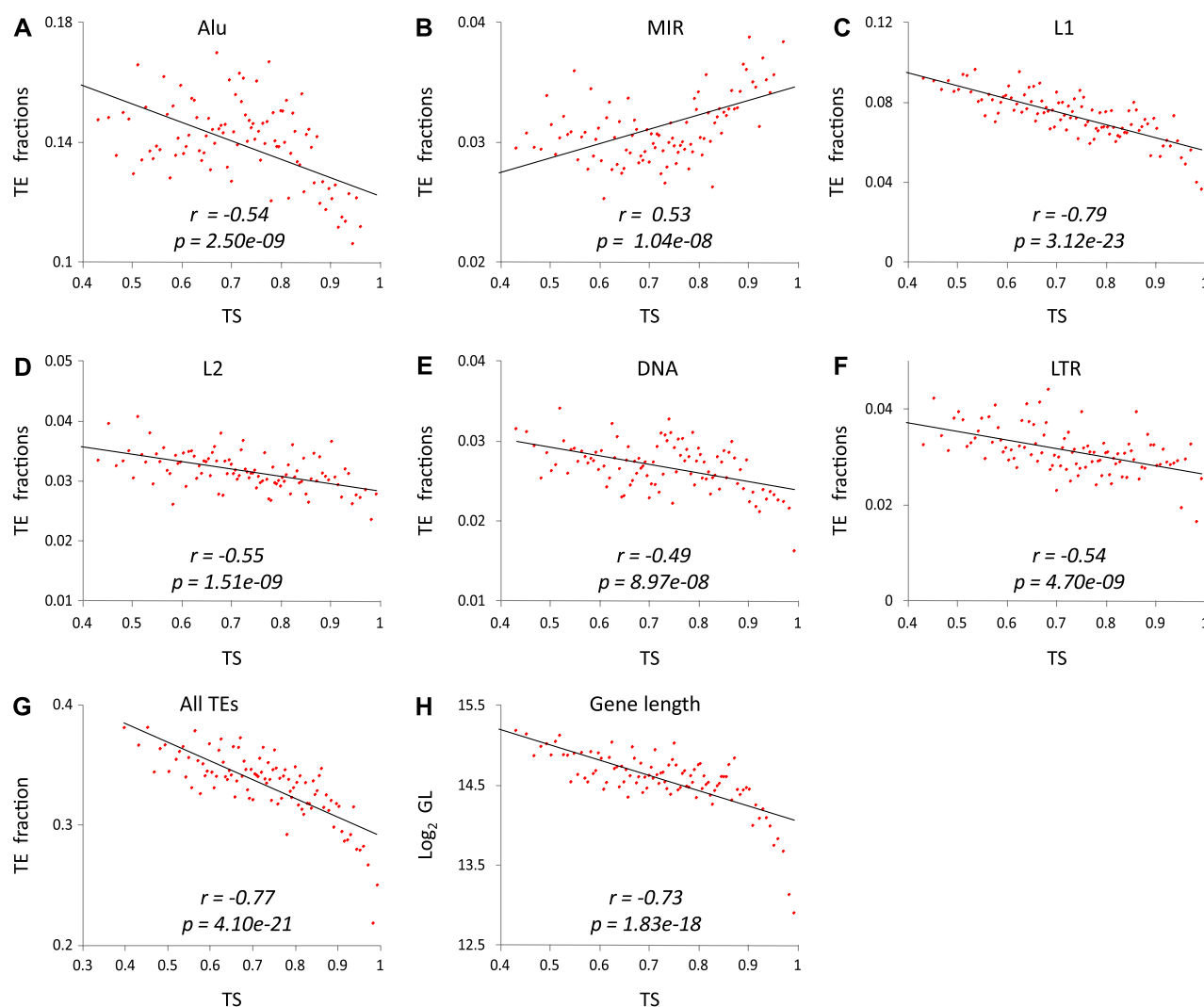


FIG. 5.—TE fractions, GL, and TS. Relationships between the TE gene fractions for (A) Alu, (B) MIR, (C) L1, (D) L2, (E) DNA, (F) LTR, and (G) all TEs and the TS of human genes. (H) Relationship between GL and TS. Pearson correlation coefficient values (r) along with their significance values (P) are shown for all pairwise regressions.

enhancer sequences (Marino-Ramirez and Jordan 2006), microRNAs (Piriyapongsa et al. 2007), and cis-natural antisense transcripts (Conley et al. 2008) to the human genome. In addition, it has been shown that, whereas TEs are generally depleted from introns, MIRs are actually significantly enriched within genes that might require subtle regulation of transcript levels or precise activation timing, such as growth factors, cytokines, hormones, and genes involved in the immune response (Sironi et al. 2006). Such genes would be expected to be largely tissue specific.

If MIRs donate regulatory sequences to tissue-specific genes, then one may expect to observe relative increases in MIR density in the regulatory regions upstream and downstream of TSSs. To evaluate this possibility, we took the top 10% tissue-specific genes and evaluated their MIR frequencies at 1-kb intervals along a 20-kb window surrounding the

gene TSS. As with all other TEs, MIRs show a marked decline in frequency most proximal to the TSS. However, MIRs show a unique pattern of enrichment both upstream and

Table 4

Effect of GC Content on the Relationship between L1 Genic Fractions and Gene Expression

	Feature ^a	r	P Value	Control ^b	r	P Value
PE	L1	−0.87	1.69E−31	L1 GC	−0.73	1.3E−25
	GC	0.69	1.20E−15	GC L1	0.12	2.2E−01
BE	L1	−0.69	1.38E−15	L1 GC	−0.44	1.7E−06
	GC	−0.21	2.00E−02	GC L1	0.44	1.4E−06
TS	L1	−0.79	3.12E−23	L1 GC	−0.77	3.0E−32
	GC	0.32	6.81E−04	GC L1	−0.03	7.5E−01

^a L1 genic fractions and genic GC content values are correlated with the expression parameters PE, BE, and TS (tissue-specificity).

^b Partial correlation analyses control for effect of GC content on L1 fractions (L1 | GC) and L1 fractions on GC content (GC | L1), respectively.

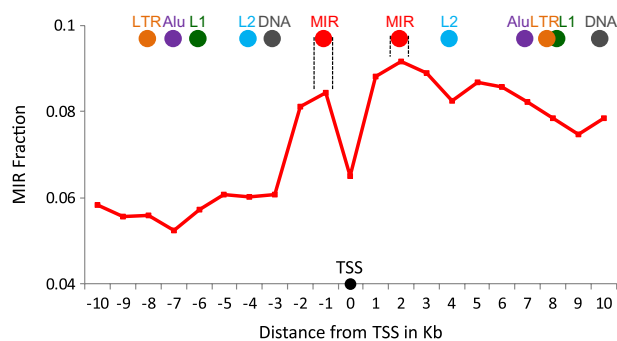


FIG. 6.—The local frequency maxima of TE densities around the TSSs of tissue-specific genes. The red line shows the density distribution of MIRs around TSSs. Colored dots show the locations of the local frequency maxima for the different TE classes/families.

downstream of the TSS, just outside the proximal promoter region, compared with other families of TEs. In fact, MIRs are the only elements that show local frequency maxima at -1 kb and $+2$ kb with respect to the TSS. All other TEs show their maxima in more distal regions from the TSS (fig. 6). This pattern is consistent with a unique regulatory role for MIRs, perhaps owing to the donation of cis-regulatory elements, as compared with other TEs.

If the regulatory effect of genic MIRs is based on the donation of shared transcription factor-binding sites, then one may expect the tissues in which MIR-rich genes are expressed to be similar. We evaluated this prediction in two ways. First, we took the top 10% MIR-rich genes and for each gene we determined the tissue in which it was maximally expressed. The observed frequency distribution for these tissues was compared with a randomized distribution of the same number of genes among all tissues in the microarray data set analyzed here using a χ^2 test. The observed distribution is far from random (supplementary fig. S4, Supplementary Material online; $\chi^2 = 1,406.8$, $P = 1.1 \times 10^{-242}$), and there are a number of specific tissues, and groups of related tissues, that are overrepresented, particularly liver, blood-related tissues, reproductive tissues and nervous tissues. Second, we clustered the expression patterns of the top 10% MIR-rich genes using hierarchical clustering based on the Euclidean distances between their gene expression

patterns over 79 tissues. Several of the resulting clusters show groups of MIR-rich genes that are markedly overexpressed among these same related groups of tissues (fig. 7).

MIRs are a relatively ancient family of TEs that are conserved among mammals including mouse. We evaluated TE gene fraction and expression data for mouse, in the same way as was done for humans, to see if the same trends in the relationship between MIR gene fractions and tissue specificity hold for mouse elements. As is the case for the human genome, mouse MIR elements are the only family of TEs with genic fractions that are significantly positively correlated with TS (table 5). This suggests the possibility that MIR elements have been conserved among mammalian genomes, at least to some extent, by virtue of their regulatory contributions.

The genomic design hypothesis predicts that additional regulatory sequence elements required by tissue-specific genes will lead to an increase in their overall length. However, with respect to MIRs, our analysis suggests that the enrichment of regulatory elements in tissue-specific genes does not lead to an increase in the overall length of genes. Rather, the regulatory complexity required by tissue-specific genes may be achieved in some cases via the donation of a few key sequence elements provided by TEs that come preequipped with existing regulatory capacity.

Conclusions

The architecture of human genes has important implications for how they are expressed. Previous studies on this topic have focused separately on the influences of GL or the TE environment on gene expression. Here, we show that these two factors are closely related, and we consider them jointly in an attempt to dissect their individual contributions. Consistent with previous results, we observed GL to be strongly correlated with PE and less so with BE. We also show that GL is strongly correlated with TS but not in the direction that is expected according to the genomic design hypothesis. These data provide strong support for the selection hypothesis. However, we show that the TE fraction of human genes has a stronger overall effect on gene expression than does GL. Considered together, TE gene fractions explain 78%,

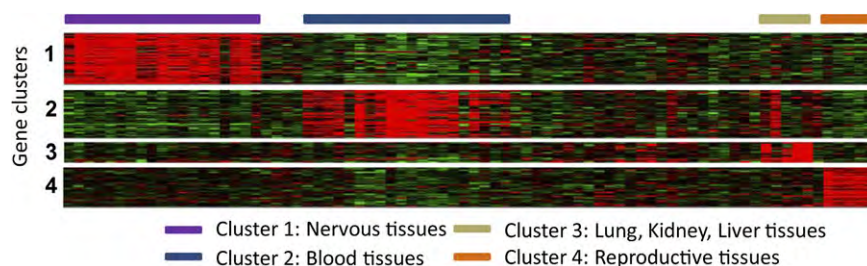


FIG. 7.—MIR-rich genes hierarchically clustered into groups of similar expression profiles across tissues. The clusters show maximum expression in related sets of tissues.

Table 5Relationship between Genic TE Fractions and Tissue-Specificity in Mouse^a

TE Family	<i>r</i>	<i>P</i> Value
MIR	0.37	7.5E-05
LTR	0.12	1.2E-01
L1	0.08	2.2E-01
DNA	0.07	2.6E-01
L2	−0.25	5.6E-03
ID	−0.40	2.1E-05
B4	−0.46	5.9E-07
B1	−0.74	1.6E-18
B2	−0.74	4.9E-19

^a Genic TE fractions for mouse TE families were correlated with tissue-specificity in the same way as done for human TE families (see fig. 5).

76%, and 66% of the variability observed for PE, BE, and TS respectively, in all cases greater than what is seen for GL. We also uncover examples where individual TE families, L1s, and MIRs respectively, have marked effects on the level and breadth of gene expression.

Consideration of intergenic TE fractions and GC content together with TE gene fractions suggests that the relationships between TE gene fractions and GL and expression are not solely related to regional genomic processes. However, there may be other as yet undetected regional genomic factors that could mitigate the apparent relationships between TE gene fractions and GL and expression. Nevertheless, the results reported here underscore the potential regulatory implications of the TE environment of human genes and also suggest specific mechanisms for how TEs may contribute to gene regulation.

Supplementary Material

Supplementary figures *S1*, *S3*, and *S4* and *table S2* are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

Acknowledgments

We would like to thank Nathan J. Bowen for guidance on the gene expression analysis. We would like to thank members of the Jordan lab for their support and technical assistance. D.J. was supported by a Fulbright predoctoral fellowship. I.K.J. and A.H. were supported by the School of Biology, Georgia Institute of Technology, and an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). This research was supported in part by the Intramural Research Program of the National Institute of Health, National Library of Medicine, NCBI.

Literature Cited

Carmel L, Koonin EV. 2009. A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. *Genome Biol Evol.* 2009:382–390.

- Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. *Nat Genet.* 31:415–418.
- Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD. 2005. Human antisense genes have unusually short introns: evidence for selection for rapid transcription. *Trends Genet.* 21:203–207.
- Comeron JM. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* 167:1293–1304.
- Conley AB, Miller WJ, Jordan IK. 2008. Human cis natural antisense transcripts initiated by transposable elements. *Trends Genet.* 24:53–56.
- Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. *Trends Genet.* 19:362–365.
- Eller CD, et al. 2007. Repetitive sequence environment distinguishes housekeeping genes. *Gene* 390:153–165.
- Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* 429:268–274.
- Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. *Gene* 345:119–126.
- Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* 110:462–467.
- Jurka J, Zietkiewicz E, Labuda D. 1995. Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Res.* 23:170–175.
- Karolchik D, et al. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32:D493–D496.
- Kim TM, Jung YC, Rhyu MG. 2004. Alu and L1 retroelements are correlated with the tissue extent and peak rate of gene expression, respectively. *J Korean Med Sci.* 19:783–792.
- Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics.* 7:474.
- Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
- Lerat E, Semon M. 2007. Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues. *Gene* 396:303–311.
- Li SW, Feng L, Niu DK. 2007. Selection for the miniaturization of highly expressed genes. *Biochem Biophys Res Commun.* 360:586–592.
- Marino-Ramirez L, Jordan IK. 2006. Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol Direct.* 1:20.
- Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res.* 12:1483–1495.
- Pereira V, Enard D, Eyre-Walker A. 2009. The effect of transposable element insertions on gene expression evolution in rodents. *PLoS One.* 4:e4321.
- Piriyapongsa J, Marino-Ramirez L, Jordan IK. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* 176:1323–1337.
- Polavarapu N, Marino-Ramirez L, Landsman D, McDonald JF, Jordan IK. 2008. Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. *BMC Genomics.* 9:226.
- Rhead B, et al. 2010. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.* 38:D613–D619.
- Seoighe C, Gehring C, Hurst LD. 2005. Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction. *PLoS Genet.* 1:e13.
- Silva JC, Shabalina SA, Harris DG, Spouge JL, Kondrashov AS. 2003. Conserved fragments of transposable elements in intergenic regions:

- evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet Res.* 82:1–18.
- Simons C, Pheasant M, Makunin IV, Mattick JS. 2006. Transposon-free regions in mammalian genomes. *Genome Res.* 16:164–172.
- Sironi M, et al. 2006. Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns. *Genome Biol.* 7:R120.
- Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev.* 9:657–663.
- Smit AF, Riggs AD. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.* 23:98–102.
- Stalteri MA, Harrison AP. 2007. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics.* 8:13.
- Sturn A, Quackenbush J, Trajanoski Z. 2002. Genesis: cluster analysis of microarray data. *Bioinformatics* 18:207–208.
- Su AI, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A.* 101:6062–6067.
- Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. *Genome Res.* 13:2260–2264.
- Ustyugova SV, Lebedev YB, Sverdlov ED. 2006. Long L1 insertions in human gene introns specifically reduce the content of corresponding primary transcripts. *Genetica* 128:261–272.
- Venter JC, et al. 2001. The sequence of the human genome. *Science* 291:1304–1351.
- Versteeg R, et al. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC-content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* 13:1998–2004.
- Vinogradov AE. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 20:248–253.
- Vinogradov AE. 2005. Dualism of gene GC-content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. *Trends Genet.* 21:639–643.
- Wang J, Bowen NJ, Marino-Ramirez L, Jordan IK. 2009. A c-Myc regulatory subnetwork from human transposable element sequences. *Mol Biosyst.* 5:1831–1839.
- Yanai I, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21:650–659.

Associate editor: Marta Wayne

Polymorphism of the Pv200L Fragment of Merozoite Surface Protein-1 of *Plasmodium vivax* in Clinical Isolates from the Pacific Coast of Colombia

Augusto Valderrama-Aguirre, Evelin Zúñiga-Soto, Leonardo Mariño-Ramírez, Luz Ángela Moreno, Ananías A. Escalante, Myriam Arévalo-Herrera, and Sócrates Herrera*

Instituto de Inmunología del Valle, Universidad del Valle, Cali, Colombia; Malaria Vaccine and Drug Development Center, Cali, Colombia; National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland; School of Life Sciences, Arizona State University, Tempe, Arizona

Abstract. Merozoite surface protein 1 (MSP-1) is a polymorphic malaria protein with functional domains involved in parasite erythrocyte interaction. *Plasmodium vivax* MSP-1 has a fragment (Pv200L) that has been identified as a potential subunit vaccine because it is highly immunogenic and induces partial protection against infectious parasite challenge in vaccinated monkeys. To determine the extent of genetic polymorphism and its effect on the translated protein, we sequenced the Pv200L coding region from isolates of 26 *P. vivax*-infected patients in a malaria-endemic area of Colombia. The extent of nucleotide diversity (π) in these isolates (0.061 ± 0.004) was significantly lower ($P \leq 0.001$) than that observed in Thai and Brazilian isolates; 0.083 ± 0.006 and 0.090 ± 0.006 , respectively. We found two new alleles and several previously unidentified dimorphic substitutions and significant size polymorphism. The presence of highly conserved blocks in this fragment has important implications for the development of Pv200L as a subunit vaccine candidate.

INTRODUCTION

Plasmodium vivax is responsible for about 20% of the global malaria cases and more than half (56%) of the non-African malarial infections. Indeed, *P. vivax* has reemerged in many regions of the world where malaria was eliminated in the 1950–60s with 70–80 million cases per year and 2.6 billion people at risk of infection.^{1,2} Given its broad distribution, *P. vivax* co-exists with *Plasmodium falciparum* and in minor proportion with *P. malariae*.^{3,4} Regardless, the progress achieved in the development of a vaccine against *P. falciparum*, there is paucity of suitable vaccine candidates against *P. vivax*, with only two antigens currently under evaluation in human clinical trials and a few others in preclinical evaluation.^{2–6} The understandable bias in research effort toward *P. falciparum*, however, hampers our ability of developing a vaccine that can be effectively deployed against malaria outside Africa where these two parasites coexist.

We previously defined the Pv200L fragment of PvMSP-1, located toward the N-terminal end of the 83 kDa domain (Figure 1), as a *P. vivax* potential subunit vaccine based on several features of the protein. The Pv200L has significant homology to Pf190L, a well-defined *P. falciparum* vaccine candidate,^{7–9} and a Pv200L recombinant protein produced in *Escherichia coli*, displayed a high level of antigenicity in humans. Additionally, this protein fragment showed good immunogenicity in mice and primates, and the capacity to induce partial protection against a *P. vivax* blood-stage challenge in *Aotus* monkeys.⁷

The Pv200L fragment includes the entire blocks 2, 3, and 4, plus segments of blocks 1 and 5 of the PvMSP-1.¹⁰ Genetic polymorphism studies in *P. vivax* isolates have shown that blocks 1, 3, and 5 are conserved at the protein level and display a few dimorphic substitutions. Blocks 2 and 4 are the most variable, both in size and sequence, with basic and recombinant block types generated by intra- and inter-allelic recombination events.¹¹ Here, we describe the polymorphism of the Pv200L

gene fragment and its inferred amino acid sequence in 26 *P. vivax* clinical isolates from Buenaventura, a malaria-endemic area located on the southern Pacific coast of Colombia.⁴ We also examined 42 Pv200L fragments from PvMSP-1 sequences previously reported in GenBank and describe the phylogenetic analysis of all sequences.

MATERIALS AND METHODS

Study population and *P. vivax* isolates. The Colombian Pacific region is composed of four states—Chocó, Valle, Cauca, and Nariño—and is considered the second most malaria-endemic region of Colombia, as it accounts for about 30% of the country's disease burden.⁴ In Colombia *P. vivax* is the predominant malaria species and is responsible for more than 60% of the clinical cases reported every year; however, because of the high prevalence of Duffy-negative in the Afro-colombian habitants of this region, the predominant species in the Pacific region is *P. falciparum*.⁴ Blood samples were collected by convenience during 2004 and 2005 from symptomatic patients diagnosed by thick smear at the outpatient clinical facilities of the Malaria Vaccine and Drug Development Center (MVDC) in Buenaventura, the main seaport on the Colombian Pacific coast. Buenaventura, located in the Valle state, has ~400,000 inhabitants and conditions for low and unstable malaria transmission.¹² A total of 26 samples exclusively infected by *P. vivax*, as confirmed by polymerase chain reaction (PCR),¹³ were collected from patients from rural and urban communities. All adult participants and the parents, or legal guardians of minor patients, were asked to provide a signed written informed consent previously approved by the Institutional Review Board of Universidad del Valle.

PCR, cloning, and sequencing. The primers 200L-1 (5'-GCC AAG CTG GAC AAG TTA GA-3') and 200L-2 (5'-AAG GTT GGA ACT GTC TTT CC-3') were designed to amplify by PCR the Pv200L coding region (bp 143–1,333) from the PvMSP-1 of Salvador 1 strain (GenBank accession no. AF435593). The primers were confirmed to align with all available complete sequences of PvMSP-1 in GenBank.¹⁴ The PCR reactions were performed with iProof high-fidelity DNA polymerase (Bio-Rad, Hercules, CA) in a PTC-100 thermal cycler (MJ Research, Watertown, MA) as follows: 35 cycles of

*Address correspondence to Sócrates Herrera, Malaria Vaccine and Drug Development Center, Carrera 37 - 2Bis No. 5E - 08, Cali, Colombia. E-mail: sherrera@inmuno.org

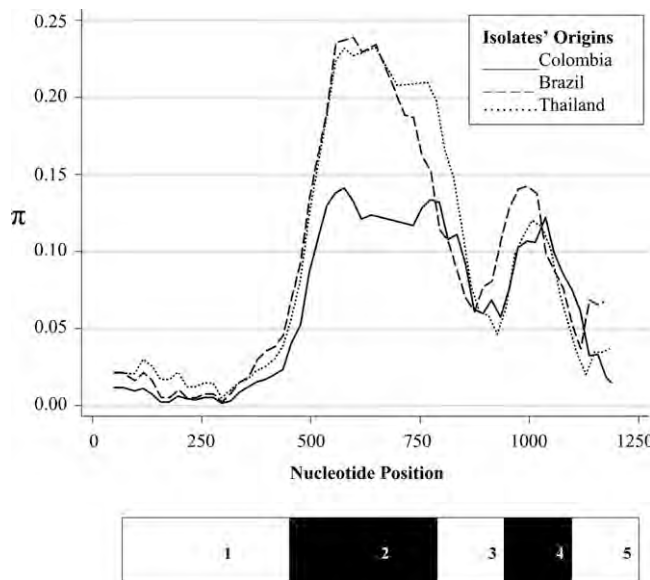


FIGURE 1. Nucleotide diversity (π) scores along the *Pv*200L gene fragment. The scores are shown for three groups of isolates: 26 from Colombia, 20 from Thailand, and 8 from Brazil. The sequences from Thai and Brazil isolates were extracted from complete *Pv*MSP-1 cds sequences available in GenBank. Relative position of blocks is indicated in the bar. Variable blocks are indicated in black.

30 sec at 94°C, 60 sec at 55°C, and 60 sec at 72°C. The PCR products (1–1.2 Kb) were ligated to pCR4-TOPO (Invitrogen, Carlsbad, CA) according to manufacturer instructions, and then used to transform chemically competent *E. coli* One-Shot (Invitrogen). Kanamycin-resistant clones were confirmed by restriction enzyme analysis with *Eco*RI (Fermentas, Hanover, MD). We sequenced only one clone per isolate. Sequencing was performed with BigDye Terminator version 3.1 kit and the M14 forward and reverse primers in an ABI-PRISM AVANT 3100 sequencer (Applied Biosystems, Foster City, CA). Every run was performed a minimum of two times and repeated as much as needed to obtain quality values > 20 (< 4 Ns/20 bases, $< 10\%$ Ns, and maximum percent of mixed bases = 20.0%) as assessed with SeqScape Software (Applied Biosystems).

Sequences, sequence alignment, and statistical analysis. We analyzed the 26 *Pv*200L gene sequences from the Colombian Pacific coast, and the 42 *Pv*200L fragments from *Pv*MSP-1 sequences available in GenBank, accession numbers: AF435593–AF435599, AF435601–AF435620, AF435622–AF435625, AF435627, AF435629–AF435632, and AF435634–AF435639 (October 2006).¹⁰ These GenBank sequences were collectively named as non-Colombian. Sequences were aligned using CLUSTAL X2,¹⁵ and the alignment was used to calculate nucleotide diversity ($\pi \pm SE$) with MEGA 4.0 using the Jukes-Cantor model and retaining all gaps.¹⁶ The π value obtained for *Pv*200L and each block was submitted to one-way analysis of variance statistical analysis. DNA polymorphism was analyzed with DnaSP 4.10 with a sliding window of 100 bases and a step size of 20 bases for a haploid genome.¹⁷ To determine the influence of geographical origin of the isolates on nucleotide diversity and DNA polymorphism, we performed the same analysis for groups of GenBank sequences from Thailand (20 isolates) and Brazil (8 isolates). The output data were exported to STATA 8.0 (Stata Corp., College Station, TX) to plot overlapping π curves with standardized scales for the

different groups. Boundaries of conserved and variable blocks were determined from nucleotide sequence homology, as previously reported.¹⁰ Subsequently, π and its standard error were computed for each block. The number of synonymous (d_s) and non-synonymous (d_n) substitutions was estimated for each block, to avoid bias caused by size polymorphisms, by Nei and Gojobori's method with the Jukes-Cantor correction in MEGA 4.0.^{18–20} We estimated the difference between D_s and D_n and its standard deviation was calculated using bootstrap with 500 pseudo-replications for D_s and D_n , and two-tail Z-test on the difference between D_s and D_n .²¹ The null hypothesis is that $D_s = D_n$; thus, we assumed as null hypothesis that the observed polymorphism was neutral. *In silico* translated DNA sequences were used to identify the types of variable blocks (2 and 4), as previously defined,¹⁰ and point mutations in conserved blocks. Alleles were identified by the specific combination of variable block types as proposed by Putaporntip and others.¹⁰

Phylogenetics and epitope conservation analysis. The distances of the *P. vivax* isolates were inferred from phylogenetic analysis of the translated *Pv*200L sequences using MEGA 4.0. Because of size polymorphism in variable blocks, distances were calculated only with sequence alignments of conserved edited-joined blocks 1, 3, and 5. Trees were constructed using the neighbor-joining (NJ) method, excluding gaps by pairwise deletion, with the Kimura p-distance model.²² The reliability of the trees was assessed by the bootstrap method with 1,000 pseudo-replications. Phylogenetic analysis was performed with the 26 Colombian sequences and the 42 sequences available in GenBank. Complementary analyses to determine genetic divergences was performed using the Fstat program (Fst). Finally, we determined the conservation pattern of previously defined promiscuous T-helper epitopes contained in the *Pv*200L fragment.²³ To this purpose we constructed multiple sequence alignments with the 68 *Pv*200L available sequences and then realigned the T-helper epitope sequences in CLUSTAL X2.

RESULTS

Geographical origin of Colombian isolates. The *Pv*200L gene fragment was sequenced and analyzed in a total of 26 isolates from the Colombian Pacific coast obtained from *P. vivax*-infected patients. Eighteen (69.2%) of the isolates were collected from patients who preferred to live in rural areas of Buenaventura (Table 1). The majority of rural isolates were from La Delfina and San Cipriano villages. The urban isolates were predominantly from Commune 12, of Buenaventura

TABLE 1
Origin of the Colombian *Plasmodium vivax* isolates

Place	Locality	n	Isolates
Comuna 12	U*	6	CU45, CU57, CU65, CU66, CU81, CU83
La Delfina	R†	6	D20, D48, D61, D85, D102, D103
San Cipriano	R	4	SC84, SC92, SC93, SC1
Buenaventura	U	2	B74, B30
La Laguna	R	1	L8
La Gloria	R	1	G91
Cordoba	R	1	Co68
Zacarias	R	1	Z80
San Francisco	R	1	SF56
Triana	R	1	T28
Citronela	R	1	Ci22
Potodo	R	1	P54

* Urban.

† Rural.

TABLE 2
Nucleotide diversity ($\pi \pm \text{SE}$)* per blocks of the *Pv*200L gene fragment

Groups	n	Overall	Block 1	Block 2	Block 3	Block 4	Block 5
All	68	0.088 \pm 0.006	0.020 \pm 0.004	0.233 \pm 0.020	0.049 \pm 0.017	0.160 \pm 0.021	0.049 \pm 0.010
Non-Colombian†	42	0.086 \pm 0.005	0.022 \pm 0.004	0.212 \pm 0.018	0.052 \pm 0.017	0.162 \pm 0.024	0.051 \pm 0.011
Colombian	26	0.061 \pm 0.004‡	0.011 \pm 0.002‡	0.155 \pm 0.014‡	0.040 \pm 0.014	0.134 \pm 0.022	0.039 \pm 0.010
Thailand	20	0.083 \pm 0.006	0.022 \pm 0.004	0.228 \pm 0.020	0.042 \pm 0.015	0.126 \pm 0.019	0.038 \pm 0.008
Brazil	8	0.090 \pm 0.006	0.021 \pm 0.004	0.203 \pm 0.019	0.056 \pm 0.020	0.166 \pm 0.026	0.035 \pm 0.015

*As calculated with MEGA 3.0 using the Jukes-Cantor model.

† *Pv* 200L gene fragment trimmed out from the 42 *Pv* MSP-1 cds available in GenBank by June 2009.

‡ $P \leq 0.001$ as calculated by one-way analysis of variance (ANOVA) Colombia vs. Thailand and Colombia vs. Brazil.

town. This commune is the closest community to the humid rain forests in the Pacific coast.

Nucleotide diversity. We found a nucleotide diversity value of 0.088 ± 0.006 for the *Pv*200L gene fragment for all the isolates, including Colombian and non-Colombian (Table 2). Such π value was significantly lower in the Colombian group than in either Thai or Brazilian subgroups of sequences. The highest contribution for nucleotide diversity was observed in the Brazilian group followed by the Thai group. As expected, the variable blocks 2 and 4 had the highest π values across all of the groups. Block 2 from Colombian isolates had a significantly lower π value than Thai and Brazilian isolates (Figure 1 and Table 2); whereas Thai sequences displayed the lowest π value for block 4. Regarding conserved blocks 1, 3, and 5, all of them displayed the expected conservation. In the case of block 1, the nucleotide diversity was significantly lower in the Colombian isolates than in the Thai and Brazilian ones. Although block 3 had also the lowest π value in the Colombian parasites, the difference was not statistically significant. Conserved block 5 displayed π values below 0.040 in all isolates, with the lowest diversity in the Brazilian isolates (Table 2).

TABLE 3

Synonymous (d_s) and non-synonymous (d_n) nucleotide substitutions in *Pv*200L fragment

Groups	Block	$d_s^* \pm \text{SE}$	$d_n^* \pm \text{SE}$	Z
All ($N = 68$)	1	0.053 \pm 0.016	0.012 \pm 0.003	-2.5944†
	2	0.168 \pm 0.036	0.259 \pm 0.032	-2.5930†
	3	0.165 \pm 0.084	0.020 \pm 0.011	-1.6597
	4	0.110 \pm 0.044	0.174 \pm 0.026	1.1718
	5	0.141 \pm 0.048	0.023 \pm 0.008	-2.5111†
Non-Colombian ($N = 42$)	1†	0.057 \pm 0.017	0.012 \pm 0.004	-2.6581†
	2	0.132 \pm 0.028	0.244 \pm 0.028	-2.67242†
	3	0.182 \pm 0.088	0.019 \pm 0.011	-1.7733
	4	0.088 \pm 0.038	0.184 \pm 0.030	1.891
	5	0.142 \pm 0.046	0.026 \pm 0.010	-2.4249
Colombian ($N = 26$)	1	0.030 \pm 0.009	0.006 \pm 0.002	-2.6278†
	2	0.142 \pm 0.029	0.161 \pm 0.020	-2.13556†
	3	0.116 \pm 0.066	0.020 \pm 0.010	-1.2895
	4	0.116 \pm 0.052	0.139 \pm 0.023	0.3967
	5	0.118 \pm 0.043	0.017 \pm 0.006	-2.1372†
Thailand ($N = 20$)	1	0.053 \pm 0.016	0.013 \pm 0.004	-2.49773†
	2	0.152 \pm 0.034	0.259 \pm 0.030	-2.42017†
	3	0.141 \pm 0.078	0.016 \pm 0.009	-1.6835
	4	0.079 \pm 0.034	0.139 \pm 0.022	1.4794
	5	0.105 \pm 0.034	0.019 \pm 0.008	-2.3153†
Brazil ($N = 8$)	1	0.060 \pm 0.018	0.010 \pm 0.003	-2.7435†
	2	0.140 \pm 0.030	0.229 \pm 0.028	-2.5384†
	3	0.184 \pm 0.090	0.024 \pm 0.013	-1.5954
	4	0.108 \pm 0.043	0.183 \pm 0.033	1.3086
	5	0.190 \pm 0.064	0.033 \pm 0.013	-2.4627†

* d_s and d_n are per every 100 sites.

† Indicates significant with $P < 0.05$.

Synonymous and non-synonymous substitutions. As expected, the d_n value was higher in blocks 2 and 4 (Table 3); however, the value was only significant in block 2. This was true for each country and the combined sample. The d_s value was significantly higher than d_n for blocks 1 and 5 using the Z-test, indicating that some blocks involved in this fragment could be under negative selection.

Non-synonymous substitutions in conserved blocks of Colombian isolates. We found a total of 26 dimorphic substitutions across conserved blocks 1, 3, and 5, and seven of them are newly identified (Table 4). Some substitutions were consistently linked within isolates, such as SC84, SC93, SC1, and Co68, most of them collected from inhabitants from San Cipriano village, which shared four dimorphic substitutions, and isolates D61 and B74 that shared five dimorphic substitutions plus two insertions (block 1: 61[–/A] and 62[–/S]).

TABLE 4

Dimorphic substitutions in conserved blocks

Block	Mutation*	Pos†	Proportion‡	Isolates§
1	N/S	53	24:02:00	D61, B74
	K/Q	56	24:02:00	D61, B74
	V/G¶	57	25:01:00	B74
	D/E	58	21:05	D61, B74, SC84, SC93, SC1
	A/T	59	24:02:00	D61, B74
	G/S	83	24:02:00	D61, B74
	S/F¶	85	25:01:00	SC1
	F/V	105	24:02:00	D61, B74
	N/H¶	108	25:01:00	SC1
	H/R¶	152	25:01:00	CU45
3	I/V¶	155	25:01:00	D48
	T/S	156	19:07	<u>Co68, SC84, SC93, SC1</u> , D61, B74, SC92
	E/D	157	22:04	Co68, SC84, SC93, SC1
	A/T	186	24:02:00	Co68, SC92
	E/A	315	17:09	L8, D20, CU45, D48, D61, CU66, B74, G91, D103
	D/G	329	22:04	<u>Co68, SC84, SC93, SC1</u>
	V/A	332	21:05	<u>Co68, SC84, SC93, SC1</u> , SF56
5	D/Y¶	384	25:01:00	SC92
	D/N	389	13:13	Ci22, T28, B30, P54, SF56, CU57, CU65, Z80, CU81, CU83, D85, SC92, D102
	I/V	409	25:01:00	SC92
	S/N¶	412	25:01:00	D103
	K/S	414	25:01:00	SC92
	S/A	415	25:01:00	SC92
	A/S	416	25:01:00	SC92
	G/S	417	25:01:00	SC92
	P/T	419	25:01:00	SC92

*The least frequent amino acid residue is presented as denominator.

† Position. Referred as to the sequence of *Pv* MSP-1 of Salvador I strain.

‡ Exact number of Colombian isolates that presented the most frequent over the least frequent amino acid residue.

§ Only those isolates that contain the least frequent substitution are listed. Isolates with linked substitutions are shown with special characters (bold, italic, and italic/bold/underlined).

¶ Newly identified substitutions.

TABLE 5
Classification of variable blocks

Type	# (%)	Isolates
Block 2a		
Basic 1	1 (3.8)	SC92
Basic 2	4 (15.4)	Co68, SC84, SC93, SC1
Basic 3	2 (7.7)	D61, B74
Basic 4	7 (26.9)	L8, D20, CU45, D48, CU66, G91, D103
Rec 6	10 (38.5)	Ci22, T28, B30, P54, SF56, CU57, Z80, CU81, CU83, D102
Rec 7	2 (7.7)	CU65, D85
Block 2b		
Type 20*	1 (3.8)	SC92
Type 14	7 (26.9)	L8, D20, CU45, D48, CU66, G91, D103
Type 15	12 (46.2)	Ci22, T28, B30, P54, SF56, CU57, CU65, Z80, CU81, CU83, D85, D102
Type 16	2 (7.7)	D61, B74
Type 19	4 (15.4)	Co68, SC84, SC93, SC1
Block 2c		
Basic 2	1 (3.8)	Co68
Basic 3	19 (73.1)	L8, † D20, † CU45, † D48, † CU66, † G91, † D103, † Ci22, T28, B30, P54, SF56, CU57, CU65, Z80, CU81, CU83, D85, D102
Rec 5	2 (7.7)	D61, B74
Rec 7	1 (3.8)	SC92
Rec 10	3 (11.6)	SC84, SC93, SC1
Block 4		
Basic 1	7 (26.9)	L8, D20, CU45, D48, CU66, G91, D103
Basic 2	4 (15.4)	Co68, SC84, SC93, SC1
Basic 3	1 (3.8)	SC92
Rec 5	14 (53.9)	Ci22, T28, B30, P54, SF56, CU57, CU65, Z80, CU81, CU83, D85, D102, D61, ‡ B74‡

= Number of block types.

* Newly identified insertion of GSSNS after fifth amino acid.

† Seven isolates had the P/S mutation in the fourth amino acid.

‡ Two isolates had the four linked substitutions I/V, D/E, V/E, and D/A at positions 3, 33, 39, and 40.

Characteristically, only isolate SC92 had seven out of nine of the dimorphic substitutions identified for block 5 (Table 4).

Variable blocks classification in Colombian isolates. Results were consistent with previously reported data.¹⁰ We identified six block 2a types (four basic and two recombinants), four block 2b types (three previously reported and a new one), five block 2c types (two basic and three recombinants), and four block 4 types (three basic and one recombinant) (Table 5). The new block 2b type named by us as type 20, had an extra insertion of the tandem repeat GSSNS in the SC92 isolate. Additionally, seven isolates shared the P/S substitution in block 2c type 3. These last two findings helped to identify the two new alleles reported here (see below). Several new substitutions were identified in each block. Interestingly, the recombinant types

identified contained fragments from basic types previously reported but not detected in the 26 Colombian isolates (see supplemental data).

Allelic distribution. The variable blocks included in the *Pv*200L fragment allowed us to classify the isolates in 12 potential alleles pooled in seven groups. Alleles are presented in groups because the definite allele needs of other variable blocks are not included in the *Pv*200L fragment (Table 6). We found two non-previously described alleles, which we have designated as 32 and 33, to continue with the nomenclature proposed previously.¹⁰ Allele 32 was designed as a new allele because the block 2b type 20 and only isolate SC92 had it. The reason why allele 33 was considered as new is because its mosaic organization of variable blocks 2 (a, b, and c) and 4 have not been previously described. A total of seven of the isolates (26.9%) were classified as allele 32 and all shared the dimorphic substitution P/S in the block 2c (Table 6). The majority of isolates (38.5%) were grouped as allele 6 or 20, meanwhile the isolate SC1 that is being used to develop a vaccine candidate was the third most frequent (11%).

Phylogenetic analysis. The NJ tree, created with distances between amino acid sequences of conserved blocks, showed that the 26 Colombian isolates clustered in agreement with the allele distribution, which was determined with the specific combination of variable blocks (Figure 2A). The isolate SC1 clustered with the isolates SC84 and SC93, all of them collected from inhabitants of San Cipriano village. The NJ tree created with the 68 *Pv*200L sequences showed that there is a slight trend to cluster in agreement with the geographical origin (Figure 2B), being more evident for the isolates from Colombia (COL), South Korea (SK), and Bangladesh (BD). Brazil (B) and Thailand (T) isolates displayed a more promiscuous clustering. The isolate SC92, collected from San Cipriano village, was the most distant of Colombian isolates, and clustered fairly close to Asian sequences (Figure 2B). A complementary analyses with *Fst* among Colombian, Brazilian, and Thai isolates revealed that there is a strong geographic structure; however, the lowest divergence (*Fst* = 0.06453) was observed between Brazilian and Thailand isolates, whereas Colombian isolates were clearly divergent (*Fst* = 0.1676 with Thailand and *Fst* = 0.2437 with Brazil).

Epitope conservation analysis. At least five different T-helper epitopes have been previously defined toward the N-terminal portion of *Pv*MSP-1, and four of them are included within the *Pv*200L fragment.²³ All of them showed a high conservation pattern along the 68 *Pv*200L amino acid sequences (Table 7).

TABLE 6
Frequency of the *Pv*Msp-1 *Pv*200L alleles identified in the Colombian Pacific coast

Block 2a	Block 2b	Block 2c	Block 4	Allele*	Isolates	Frequency (%)
Rec 6	Type 15	Basic 3	Rec 5	6, 20	Ci22, T28, B30, P54, SF56, CU57, Z80, CU81, CU83, D102	38.5
Basic 4	Type 14	Basic 3 †	Basic 1	33‡	L8, D20, CU45, D48, CU66, G91, D103	26.9
Basic 2	Type 19	Rec 10	Basic 2	7, 21	SC84, SC93, SC1	11.5
Basic 3	Type 16	Rec 5	Rec 5	5, 11, 12, 13	D61, B74	7.7
Rec 7	Type 15	Basic 3	Rec 5	25	CU65, D85	7.7
Basic 1	Type 20§	Rec 7	Basic 3	32	SC92	3.8
Basic 2	Type 19	Basic 2	Basic 2	31	Co68	3.8

* Allele is defined by the specific arrangement of variable blocks. Several alleles mean that any of them contain that specific combination of blocks 2a, 2b, 2c, and 4. To define the exact allele needs variable blocks that are not included in the *Pv*200L sequence.

† P/S mutation in the fourth amino acid across the seven isolates.

‡ New alleles.

§ Newly identified insertion of repeat GSSNS after fifth amino acid.

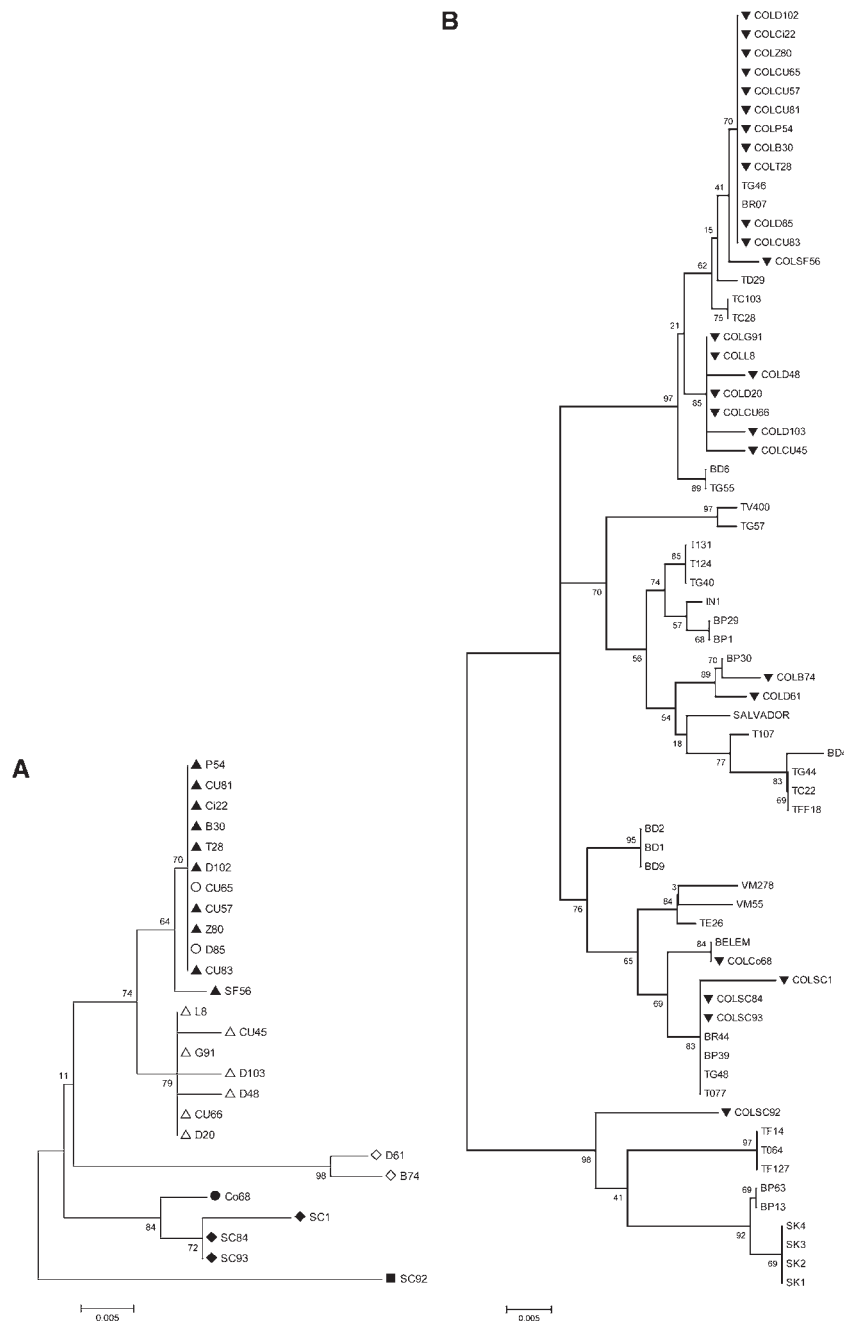


FIGURE 2. Neighbor-joining (NJ) phylogenetic trees of *Pv200L*. Trees were built with edited-joined sequences of conserved blocks 1, 3, and 5. (A) NJ tree of the 26 *Pv200L* sequences from Colombian *Plasmodium vivax* isolates. Isolates alleles are denoted as follows: ▲ (Alleles 6 or 20), △ (Allele 33), ◆ (Alleles 7 or 21), ◇ (Alleles 8, 11, 12 or 13), ● (Allele 31), and ■ (Allele 32). (B) NJ trees of trees of 68 *Pv200L* sequences, including 42 obtained from *PvMSP-1* cds available in Genbank and the 26 Colombian isolates (▼COL).

DISCUSSION

As expected, the *Pv200L* fragment of *PvMSP-1* showed to be polymorphic in sequence and in size, which resembles very well the mosaic structure previously described by others, with a polymorphism greatly concentrated in fragments 2 and 4, surrounded by the well-conserved blocks 1, 3, and 5.¹⁰ However, regardless of this polymorphism, there is some evidence of purifying selection. Although this study focused only in the *Pv200L* fragment, our results may indicate a more complex dynamic of selection acting on the *PvMSP-1* polymorphism

than previously thought.²⁴ The overall nucleotide polymorphism and the genetic diversity found in the isolates from the Colombian Pacific coast was substantially lower than in those from Thailand and Brazil, probably due in part to the lower/seasonal transmission and geographic isolation. Most malaria cases studied here occur in close rural communities with high internal mobility but low foreign influence from potentially *P. vivax*-infected international travelers. This geographic isolation also could explain the pattern of *Fst*. However, presence of imported parasites cannot be completely excluded. SC92

TABLE 7

Conservation pattern of promiscuous T-helper epitopes in Pv200L

Id§	Sequence conservation
PvT4	NFVGKFLQLQIPGHTDLLHL *:*.******
PvT6	FNQLMHVINFHFDLLRAKLH *****
PvT8	LDMLKKVVLGLYRPLDNIKD *****
PvT19	LEYYLREKAKMAGTLIIPES *****.*:**

§As reported by Caro-Aguilar and others.²³ Asterisk stands for 100% conserved residue across 68 sequences analyzed.

was the most distant Colombian isolate; it clusters with Asian sequences, and is a non-previously described allele. These observations might indicate the introduction of PvMSP-1 alleles from Asian *P. vivax*. This possibility is not extraordinary given that Buenaventura is the main entrance for most of the Colombian market arriving through the Pacific Ocean. Inter- and intra-allelic recombination can be observed among the Colombian isolates, a factor that can be attributable to the active internal migration of *P. vivax*-infected patients and multiplicity of infection within the study area, which might be also the explanation why there is a low correlation between Colombian alleles and the geographical origin.

The distribution of Colombian isolates in the NJ tree resembled very well the allele groups identified. The former was built based exclusively in conserved blocks 1, 3, and 5; meanwhile, the allele is defined exclusively in the specific combination of variable blocks 2a, 2b, 2c, and 4. Such a specific combination of variable blocks (allele), in some cases, was associated with specific dimorphic substitution observed in conserved residues of either variable or conserved blocks. For example, 1) all isolates having the block 2c basic 3 with the substitution P/S in the fourth position were finally classified as being allele 33, meanwhile the remaining basic 3 (without the P/S) substitution were classified as alleles 6 or 20; and 2) a similar effect was observed with isolates D61 and B74, which always matched together for variable block 2a–c and shared nine exclusively linked dimorphic substitutions, five of them in conserved block 1 and four in variable block 4. Despite no clear geographical relationship, the similarity between them is so strong that it is possible to propose that the isolate from Buenaventura (B74), where malaria transmission does not occur, was acquired in La Delfina (D61), which is a near rural spot for local tourism with active malaria transmission.

Several non-previously described dimorphic substitutions and two new alleles were found suggesting that Pv200L from PvMSP-1 is under evolutionary forces non-equally distributed along the whole fragment. Such evolutionary forces, which might be related to the immune response, may have generated these new alleles by selecting new combinations generated by mitotic recombination in the asexual blood stages.

Besides the presence of highly conserved fragments in this protein, several other factors indicate that this protein could be a suitable target for a vaccine: First, this protein is highly antigenic as indicated by its recognition by the great majority of individuals from endemic communities in Brazil and Colombia.⁷ Second, it has been demonstrated that naturally acquired IgG antibodies directed to the N-terminal region of PvMSP-1 are associated with clinical protection to *P. vivax*-malaria.²⁵ Third, previously defined promiscuous T-helper

epitopes are located at highly conserved portions of Pv200L and displayed a high conservation pattern across the 68 sequences of Pv200L.

This is the first study of Pv200L polymorphism in Colombia. Although the number of isolates is limited, this is the PvMSP-1 polymorphism study with the highest number of *P. vivax* isolates from the same geographic origin. Such an approach has allowed us to confirm the mosaic structure of PvMSP-1 and its inter- and intra-allelic recombinant nature, to illustrate its specific association with dimorphic substitutions in conserved blocks, and describe the wide geographic distribution of highly conserved epitopes, despite the high level of polymorphism in this fragment. Further studies with a larger number of isolates from other endemic regions of the country and the planet would be required to assess genetic diversity with greater accuracy, linkage disequilibrium, and population structure. The presence of highly conserved blocks in this fragment of the *P. vivax* MSP-1 protein has important implications for the development of Pv200L as a subunit vaccine candidate.

Received September 2, 2009. Accepted for publication December 23, 2009.

Acknowledgments: We thank Ingrid Felger from the Swiss Tropical Institute (Basel, Switzerland) and Hernando del Portillo from CRESIB (Barcelona, Spain) for their advice and criticism, and Suzanne Fischer from Family Health Organization (USA) for her assistance in revising and editing this manuscript.

Financial support: This study was supported through a *P. vivax* vaccine research program granted by the National Institute of Allergy and Infectious Diseases (NIAID grant no. A1-49486/ TMRC), National Bureau of Sciences, University of Valle State (contract no. 245-2004), COLCIENCIAS (grant 1106-04-16489), and the Colombian Ministry of Social Protection (grant 2304-04-19524) (contract no.253-2005). Ananías A. Escalante is supported by the grant R01GM080586 from the National Institutes of Health, USA and through an International Center of Excellence for Malaria Research NIAID/ICEMR grant no U 19AI089702.

Authors' addresses: Augusto Valderrama-Aguirre, Evelin Zúñiga-Soto, Luz Ángela Moreno, Myriam Arévalo-Herrera, and Sócrates Herrera, Instituto de Immunología, Facultad de Salud, Universidad del Valle, Cali, Colombia and Malaria Vaccine and Drug Development Center, Cali, Colombia, E-mails: avalderrama@inmuno.org, ezuniga@inmuno.org, luzangelmo@hotmail.com, marevalo@inmuno.org, and sherrera@inmuno.org. Leonardo Mariño-Ramírez, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Computational Biology Branch, Building 38A, Bethesda, MD, E-mail: marino@ncbi.nlm.nih.gov. Ananías A. Escalante, School of Life Sciences, Arizona State University, Tempe, AZ, E-mail: Ananias.Escalante@asu.edu.

Reprint requests: Sócrates Herrera, Malaria Vaccine and Drug Development Center, Carrera 37 - 2Bis No. 5E - 08, Cali, Colombia, E-mail: sherrera@inmuno.org.

REFERENCES

- Guerra CA, Hay SI, Lucio-Parades LS, Gikandi PW, Tatem AJ, Noor AM, Snow RW, 2007. Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project. *Malar J* 6: 17.
- Mendis K, Sina BJ, Marchesini P, Carter R, 2001. The neglected burden of *Plasmodium vivax* malaria. *Am J Trop Med Hyg* 64: 97–106.
- Korenromp E, Miller J, Nahlen B, Wardlaw T, Young M, 2005. World malaria report. RBM/WHO/UNICEF, ed. Geneva, Switzerland: Roll Back Malaria, 1–214.
- Subdirección de Vigilancia y Control en Salud Pública INdS, 2006. Informe del VI período epidemiológico: malaria. SIVIGILA Semanas 1–24: 21–25.

5. Herrera S, Bonelo A, Perlaza BL, Fernandez OL, Victoria L, Lenis AM, Soto L, Hurtado H, Acuna LM, Velez JD, Palacios R, Chen-Mok M, Corradin G, Arévalo-Herrera M, 2005. Safety and elicitation of humoral and cellular responses in Colombian malaria-naïve volunteers by a *Plasmodium vivax* circumsporozoite protein-derived synthetic vaccine. *Am J Trop Med Hyg* 73: 3–9.
6. Polley SD, McRobert L, Sutherland CJ, 2004. Vaccination for vivax malaria: targeting the invaders. *Trends Parasitol* 20: 99–102.
7. Valderrama-Aguirre A, Quintero G, Gomez A, Castellanos A, Perez Y, Mendez F, Arévalo-Herrera M, Herrera S, 2005. Antigenicity, immunogenicity, and protective efficacy of *Plasmodium vivax* MSP1 Pv200L: a potential malaria vaccine subunit. *Am J Trop Med Hyg* 73: 16–24.
8. Genton B, Al-Yaman F, Betuela I, Anders RF, Saul A, Baea K, Mellombo M, Taraika J, Brown GV, Pye D, Irving DO, Felger I, Beck HP, Smith TA, Alpers MP, 2003. Safety and immunogenicity of a three-component blood-stage malaria vaccine (MSP1, MSP2, RESA) against *Plasmodium falciparum* in Papua New Guinean children. *Vaccine* 22: 30–41.
9. Guttinger M, Romagnoli P, Vandel L, Meloen R, Takacs B, Pink JR, Sinigaglia F, 1991. HLA polymorphism and T cell recognition of a conserved region of p190, a malaria vaccine candidate. *Int Immunol* 3: 899–906.
10. Putaporntip C, Jongwutiwes S, Sakihama N, Ferreira MU, Kho WG, Kaneko A, Kanbara H, Hattori T, Tanabe K, 2002. Mosaic organization and heterogeneity in frequency of allelic recombination of the *Plasmodium vivax* merozoite surface protein-1 locus. *Proc Natl Acad Sci USA* 99: 16348–16353.
11. Putaporntip C, Jongwutiwes S, Tanabe K, Thaithong S, 1997. Interallelic recombination in the merozoite surface protein 1 (MSP-1) gene of *Plasmodium vivax* from Thai isolates. *Mol Biochem Parasitol* 84: 49–56.
12. Gonzalez JM, Olano V, Vergara J, Arévalo-Herrera M, Carrasquilla G, Herrera S, Lopez JA, 1997. Unstable, low-level transmission of malaria on the Colombian Pacific Coast. *Ann Trop Med Parasitol* 91: 349–358.
13. Snounou G, 1996. Detection and identification of the four malaria parasite species infecting humans by PCR amplification. *Methods Mol Biol* 50: 263–291.
14. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL, 2007. GenBank. *Nucleic Acids Res* 35: 21–25.
15. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG, 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25: 4876–4882.
16. Kumar S, Tamura K, Nei M, 2004. MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 5: 150–163.
17. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R, 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496–2497.
18. Tajima F, 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.
19. Fu YX, Li WH, 1993. Statistical tests of neutrality of mutations. *Genetics* 133: 693–709.
20. Nei M, Gojobori T, 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3: 418–426.
21. Nei M, Kumar S, 2000. *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.
22. Saitou N, Nei M, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4: 406–425.
23. Caro-Aguilar I, Rodriguez A, Calvo-Calle JM, Guzman F, De la Vega P, Patarroyo ME, Galinski MR, Moreno A, 2002. *Plasmodium vivax* promiscuous T-helper epitopes defined and evaluated as linear peptide chimera immunogens. *Infect Immun* 70: 3479–3492.
24. Tanabe K, Escalante A, Sakihama N, Honda M, Arisue N, Horii T, Culleton R, Hayakawa T, Hashimoto T, Longacre S, Pathirana S, Handunnetti S, Kishino H, 2007. Recent independent evolution of msp1 polymorphism in *Plasmodium vivax* and related simian malaria parasites. *Mol Biochem Parasitol* 156: 74–79.
25. Nogueira PA, Alves FP, Fernandez-Becerra C, Pein O, Santos NR, Pereira da Silva LH, Camargo EP, del Portillo HA, 2006. A reduced risk of infection with *Plasmodium vivax* and clinical protection against malaria are associated with antibodies against the N terminus but not the C terminus of merozoite surface protein 1. *Infect Immun* 74: 2726–2733.

RESEARCH

Open Access

Epigenetic histone modifications of human transposable elements: genome defense versus exaptation

Ahsan Huda¹, Leonardo Mariño-Ramírez^{2,3}, I King Jordan^{1*}

Abstract

Background: Transposition is disruptive in nature and, thus, it is imperative for host genomes to evolve mechanisms that suppress the activity of transposable elements (TEs). At the same time, transposition also provides diverse sequences that can be exapted by host genomes as functional elements. These notions form the basis of two competing hypotheses pertaining to the role of epigenetic modifications of TEs in eukaryotic genomes: the genome defense hypothesis and the exaptation hypothesis. To date, all available evidence points to the genome defense hypothesis as the best explanation for the biological role of TE epigenetic modifications.

Results: We evaluated several predictions generated by the genome defense hypothesis versus the exaptation hypothesis using recently characterized epigenetic histone modification data for the human genome. To this end, we mapped chromatin immunoprecipitation sequence tags from 38 histone modifications, characterized in CD4+ T cells, to the human genome and calculated their enrichment and depletion in all families of human TEs. We found that several of these families are significantly enriched or depleted for various histone modifications, both active and repressive. The enrichment of human TE families with active histone modifications is consistent with the exaptation hypothesis and stands in contrast to previous analyses that have found mammalian TEs to be exclusively repressively modified. Comparisons between TE families revealed that older families carry more histone modifications than younger ones, another observation consistent with the exaptation hypothesis. However, data from within family analyses on the relative ages of epigenetically modified elements are consistent with both the genome defense and exaptation hypotheses. Finally, TEs located proximal to genes carry more histone modifications than the ones that are distal to genes, as may be expected if epigenetically modified TEs help to regulate the expression of nearby host genes.

Conclusions: With a few exceptions, most of our findings support the exaptation hypothesis for the role of TE epigenetic modifications when vetted against the genome defense hypothesis. The recruitment of epigenetic modifications may represent an additional mechanism by which TEs can contribute to the regulatory functions of their host genomes.

Background

Transposable elements (TEs) are mobile DNA sequences that can replicate to extremely high genomic copy numbers. TEs are also widely distributed; they have been found within genomes representing all major eukaryotic lineages. Accordingly, TEs have had a profound impact on the structure, function and evolution of their host genomes. In this study, we explore the relationship

between TEs and the epigenetic regulatory mechanisms that are thought to have evolved in response to their proliferation in eukaryotic genomes [1].

Transposition is inherently disruptive in nature. Therefore, in order to ensure their own survival, host genomes must have evolved various repressive mechanisms to guard against deleterious TE insertions. Epigenetic regulatory modifications represent a broad class of silencing mechanisms that may have come into existence in response to the need to repress TEs [1-4]. The notion that epigenetic regulatory systems evolved to

* Correspondence: king.jordan@biology.gatech.edu

¹School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30332, USA

silence TEs is known as the 'genome defense hypothesis' [4] and this hypothesis can be taken to make several predictions regarding the epigenetic modifications of TEs. According to the genome defense hypothesis, it may be expected that: (1) younger TEs, that is those that are potentially active, will bear more epigenetic modifications than older inactive TEs; and (2) TEs will bear primarily repressive (gene silencing) modifications rather than active modifications which are associated with gene expression.

An alternative hypothesis to the genome defense model is what we refer to as the 'exaptation hypothesis'. An exaptation describes an organismic feature that currently performs a function for which it was not originally evolved [5]. In the case of TEs, it is well known that a number of formerly selfish or parasitic element sequences have been exapted to provide regulatory and/or coding sequences that serve to increase the fitness of the host [6,7]. For instance, TEs can regulate host genes by serving as the targets of epigenetic histone modifications that spread into adjacent gene loci [2,8]. TE sequences that have been exapted are often anomalously conserved, due to the fact that they are preserved by natural selection after acquiring a function for the host genome [9]. For this reason, exapted TEs tend to be relatively ancient compared to TEs genome-wide.

Consideration of the exaptation hypothesis for TEs in epigenetic terms also yields several specific predictions. According to the TE exaptation model, it is expected that: (1) older and more conserved TEs will bear more epigenetic marks than younger TEs; (2) both active and repressive histone modifications will be targeted to TEs; and (3) TEs closer to genes will bear more histone modifications than more distal TEs.

Our current understanding of the relationship between TEs and epigenetic histone modifications is mainly derived from studies on plants and fungi [10-17]. The vast majority of evidence from these studies points to the genome defense hypothesis as the best explanation for how and why TEs are epigenetically modified. For instance, in *Arabidopsis thaliana*, TE insertions can trigger *de novo* formation of heterochromatin by recruiting repressive histone modifications [2,10]. Similarly, in the yeast *Schizosaccharomyces pombe*, a classical repressive histone tail modification histone H3 lysine 9 trimethylation (H3K9me3) is known to induce the formation of heterochromatin upon a TE insertion [18]. For both plants and yeast, RNA transcripts generated from TEs are thought to trigger an RNA interference related pathway that leads to their epigenetic suppression [13,14].

To date, only a handful of studies have investigated the relationship between mammalian TEs and epigenetic histone modifications. These studies have found that

mammalian TEs are targeted primarily by repressive histone tail modifications. The first indication of the involvement of repressive histone modifications with human TEs was unexpectedly discovered by Kondo and Issa in 2003 who found that H3K9me2 is targeted primarily to Alu elements in the human genome [19]. A couple of years later, Martens *et al.* reported varying levels of TE enrichment for repressive marks in repetitive DNA in mouse embryonic stem cells [20]. Recently, a genome-wide map of several histone tail modifications in mouse was published by the Bernstein and Lander groups [8,21]. They found that intracisternal A particle (IAP) and early transposon (ETn) elements were the only families of TEs enriched in repressive histone marks. IAP and ETn are young and active lineages of long terminal repeat (LTR) - retrotransposons and their targeting by repressive modifications is consistent with the host's need to suppress their activity. Another recent study in the mouse by the Jenuwein group also found an enrichment of the repressive mark H3K27me3 in silent genes and nearby short interspersed nuclear elements (SINEs) [22]. Thus, the majority of evidence to date points to the genome defense hypothesis as the best explanation for the role of epigenetic modifications targeted to mammalian TE sequences.

Recently, a series of chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) experiments have been performed by the Keji Zhao group, which together yield a genome-wide map of histone tail modifications in human CD4⁺ T cells [23,24]. These data provide a unique opportunity to qualitatively and quantitatively investigate the relationship between epigenetic histone modifications and human TEs, and to test the predictions of the genome defense hypothesis versus the exaptation hypothesis.

Results and discussion

Characterization of TE histone modifications

Previously, a series of ChIP-Seq analyses were used to determine the genome-wide distributions of 38 histone tail modifications in human CD4⁺ T cells [23,24]. For these studies, sequence tags corresponding to specifically modified histones were characterized using the Illumina-Solexa platform and the tags were mapped to the human genome sequence using the software provided by the vendor. This approach only yields unambiguously mapped sequence tags that correspond to unique genomic locations. In other words, all tags that map to repetitive sequences are eliminated from consideration. Since we are analysing TEs here, many of which are repetitive DNA sequences, we used our own mapping procedure (see Methods) to recover many of the sequence tags that map to more than one location in the genome and therefore had been discarded in the previous studies.

Our tag-to-genome mapping procedure yielded a total of 369,225,759 mapped sequence tags over the 38 histone modifications. This figure represents an increase of 144,125,239 tags (64%) over the previously employed mapping procedure, for an average increase of 3,792,769 tags per modification. Differences in the numbers of mapped tags for each histone modification can be seen in Additional file 1, Figure S1. For human TE sequences, we mapped an additional 77,065,760 tags over the 38 modifications.

The genome defense hypothesis for TE epigenetic modifications predicts that TEs will bear primarily repressive, rather than active, histone tail modifications, whereas the exaptation hypothesis holds that both active and repressive histone modifications will be targeted to TEs. The histone tail modifications analysed here were characterized as active or repressive based on their enrichment in genes with different CD4⁺ T cell expression levels using a previously described approach [24]. To apply this approach, we established presence/absence calls for each modification in the promoter regions of human genes by comparing promoter modification tag counts to corresponding genomic background tag counts as described in the Methods. We then calculated the fold enrichment of expression by comparing the average CD4⁺ T cell expression level of genes marked as present for a particular modification with the average expression level of genes that do not display any enrichment of the same modification (Additional file 1, Figure S2). There are 28 histone tail modifications characterized as active using this approach and 10 modifications characterized as repressive. This method reveals the effects of individual histone modifications on gene expression, presumably based on how they help to determine open versus closed chromatin states. In other words, active modifications are associated with the active expression of human gene sequences, whereas repressive modifications are associated with gene silencing. Accordingly, the genome defense hypothesis would predict the targeting of potentially active TEs with repressive histone tail modifications.

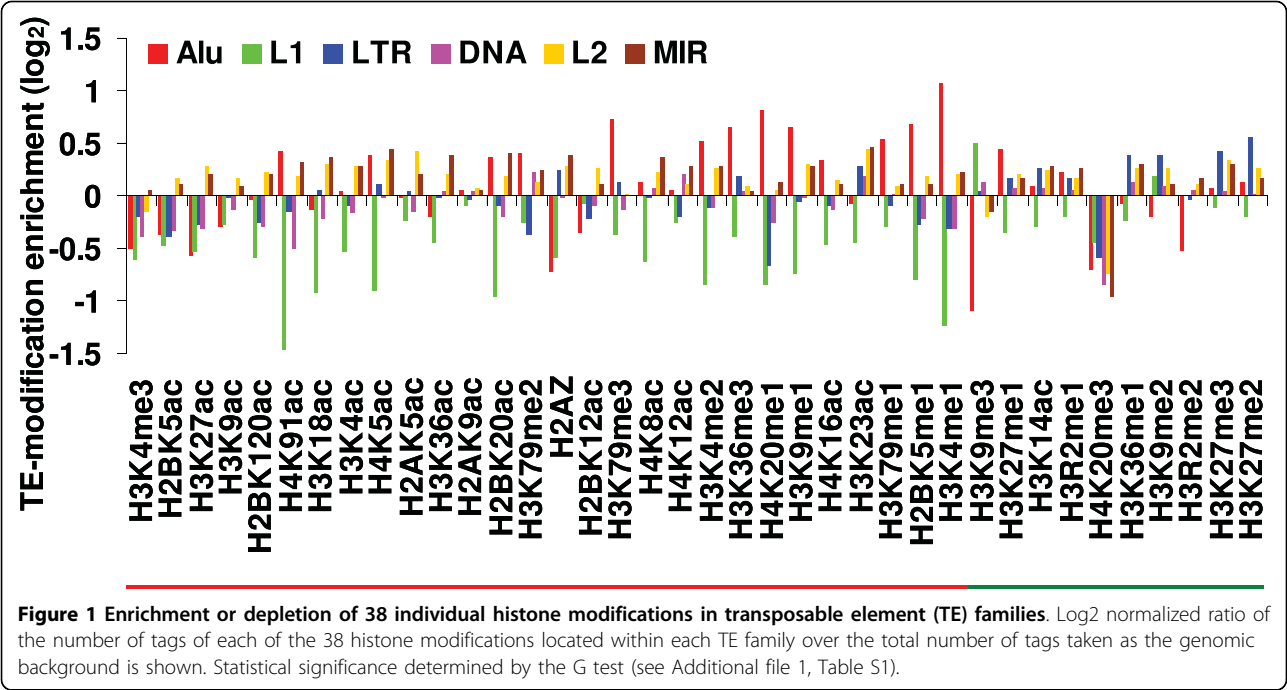
A variety of TEs are found in the human genome [25]. Retrotransposons constitute the vast majority of these sequences with Alu and L1 being the youngest and most abundant families and MIR and L2 being older inactive lineages of SINEs and LINEs, respectively. LTR retrotransposons are a less abundant but more diverse group of retrotransposons, with very few extant subfamilies. DNA-type elements make up a distinct class of TEs, which are substantially less abundant than retrotransposons in the human genome. We evaluated the relative enrichment of each histone tail modification over six classes (families) of human TEs: Alu, L1, LTR, DNA, L2 and MIR (Figure 1). To do this, a fold-change

approach similar to that used to characterize active versus repressive modifications was used. For each histone tail modification, the TE family-specific tag counts were compared against the genomic background for that modification (Methods). Thus, the fold-change values represent the extent to which TE families are enriched or depleted for each of the 38 histone tail modifications. This generated a total of 228 (6 × 38) TE-by-modification fold-change values, all of which were statistically significant (Additional file 1, Table S1; G test $0 = P < 2.1e-5$). TE epigenetic histone modifications vary widely according to the TE family, as well as the identity of the specific modification. There are numerous active and repressive modifications that are enriched for different TE families. Some families, such as Alu and L2, appear to be enriched for active modifications, whereas others, such as L1 and LTR, are depleted for active modifications and/or enriched for repressive modifications. Clearly, human TE sequences are bound by histones that are subject to numerous active and repressive epigenetic modifications.

Human TEs are distributed non-randomly across the genome with respect to gene locations and guanine-cytosine (GC) content. For instance, Alu elements are enriched in and around genes in high GC rich regions of the genome, whereas L1 elements are found primarily in AT rich DNA in intergenic regions [25]. Thus, using the entire genomic background of histone modification tag counts to compute the modification enrichments for TE families with distinct genomic distributions could bias the results. In order to control for this possibility, we re-calculated the enrichment of histone modifications by comparing the histone modification tag counts of each TE to a background tag count computed from a genomic window encompassing that TE (Methods). This local approach to computing TE histone modification enrichments does not qualitatively change the results obtained when compared to the global approach. Indeed, the TE-histone modification enrichment ratios computed using global versus local histone modification background tag counts are highly correlated ($0.91 = r = 0.99$) for each of the six classes (families) of TEs evaluated (Additional file 1, Figure S3). For comparison, the relative enrichments of TE-histone tail modifications calculated in this way are shown in Additional file 1, Figure S4. Whether the TE-histone modification enrichments are computed using global or local modification tag counts, human TEs show evidence of being targeted by a number of different active and repressive epigenetic marks.

Active versus repressive TE histone modifications

The genome defense hypothesis for TE epigenetic modifications predicts that TEs that are capable of transposition will be targeted by repressive histone modifications



in order to suppress their activity. The exaptation hypothesis, on the other hand, predicts that older and more conserved TEs will bear more epigenetic marks. These older TEs will have lost the ability to transpose and are more likely to have been exapted to play some role for their host genome. To distinguish between these models, we correlated the histone tail modification enrichment for specific TE families with the histone tail modification gene expression enrichment values. The genome defense hypothesis would predict a negative correlation since repressive modifications should target actively expressing TEs with the potential to transpose, whereas the exaptation model may predict a positive correlation or no correlation at all. None of the TE families shows a statistically significant relationship between TE and gene expression enrichment for individual histone modifications (Figure 2 and Additional file 1, Table S2). The same analysis was done using the local approach to computing the histone modification background tag counts, as described in the previous section, and the results are qualitatively similar when this technique is applied (Additional file 1, Figure S5). These results are not consistent with the genome defense hypothesis, but it is unclear whether they reflect the absence of genome defense, exaptation or some combination thereof.

To further evaluate the active versus repressive TE modification predictions for the genome defense versus exaptation hypotheses, we grouped and summed the histone tail modification tags into the 28 active and 10

repressive modifications. The enrichment of active and repressive modifications was calculated by co-locating the tags from each class with TE sequences from each family and comparing the TE family-specific active or repressive tag counts with the genomic background. The data shows considerable variation between active and repressive modification enrichments in different lineages of TEs (Figure 3). Alus and L1s are significantly depleted in both active and repressive modifications, with relatively fewer active modifications. LTR elements show depletion for active modifications and enrichment for repressive modifications, which is entirely consistent with the predictions of the genome defense model. On the other hand, L2 and mammalian-wide interspersed repeat (MIR) elements show enrichment for both active and repressive modifications consistent with the exaptation model.

The data on active versus repressive histone modifications for TE families also bears on the predictions relating epigenetic modifications to the ages of TEs. The genome defense hypothesis predicts that potentially active younger TEs will bear more epigenetic modifications than older TEs, while the exaptation model predicts that more ancient conserved TEs will bear more epigenetic modifications. The different families of TEs shown in Figure 3 have different relative ages, on average, with Alu elements being the youngest and MIRs being the oldest [young-to-old: Alu-L1-LTR-DNA-L2-MIR] [25]. The enrichments of both active and repressive modifications are positively correlated with the age

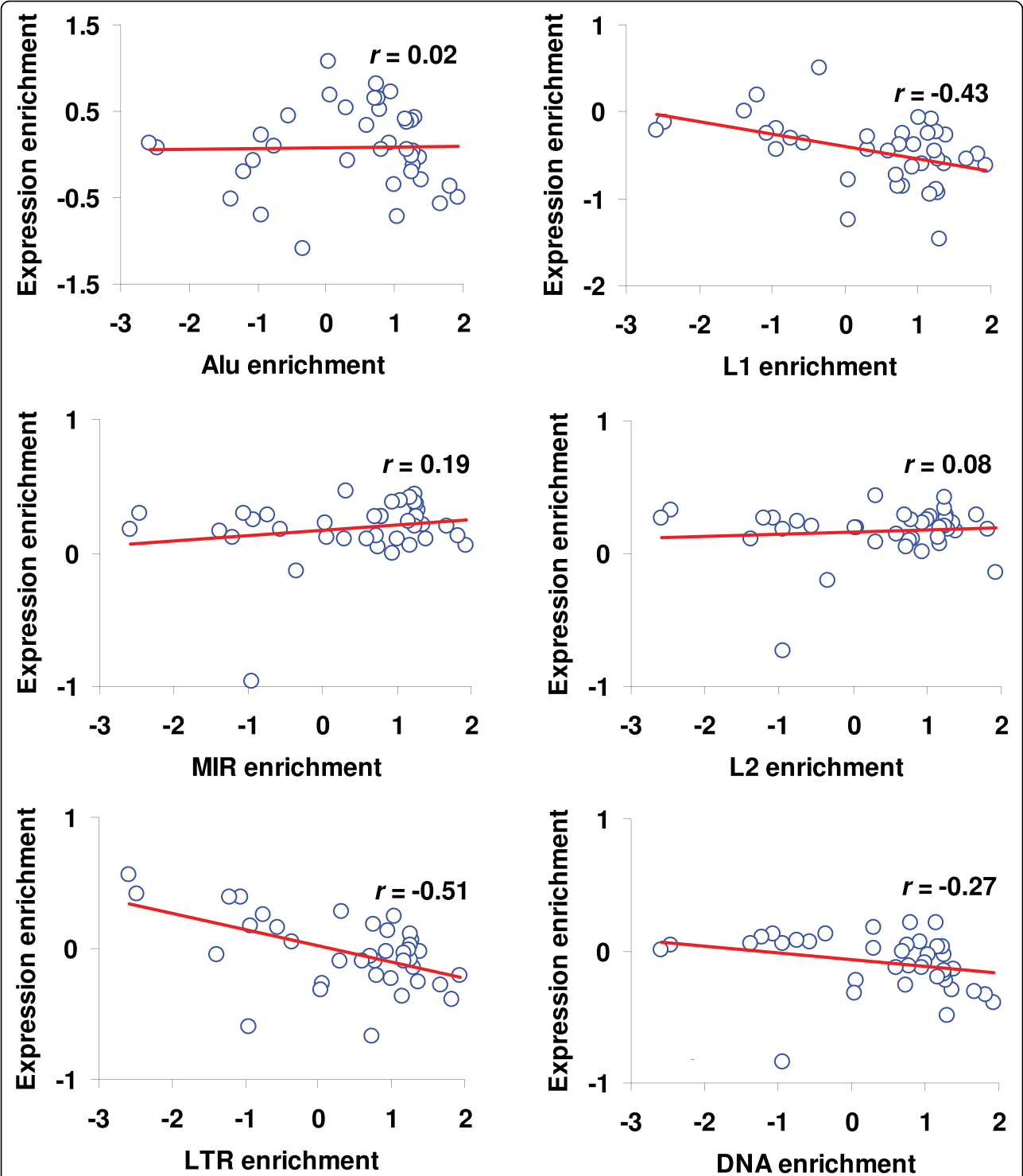
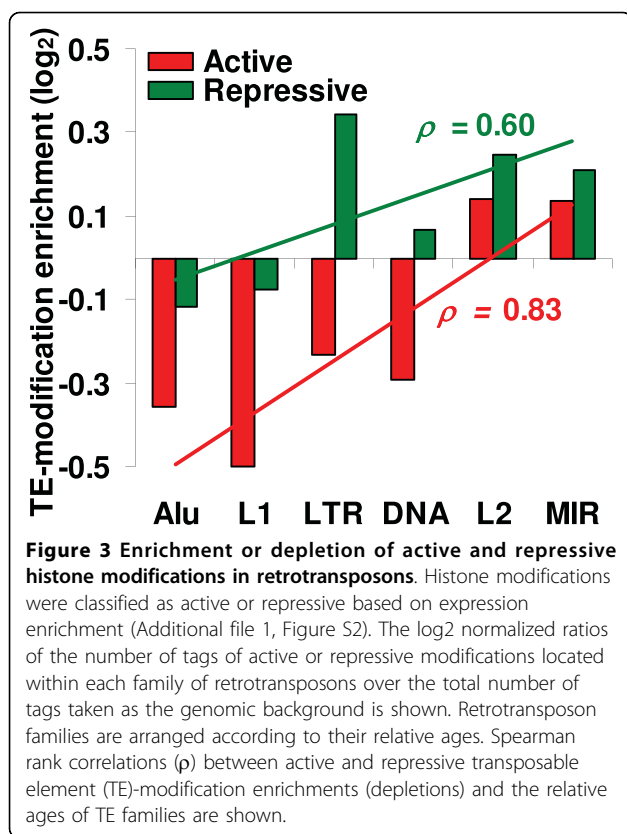


Figure 2 Correlation between enrichment of histone modifications in transposable element (TE) families and for human gene expression. The enrichment of 38 histone modifications in human gene expression (Additional file 1, Figure S2) is plotted against the same in six TE families (Figure 1). See Methods for details and Additional file 1, Table S2 for statistical significance. Pearson correlation coefficient values (r) are shown.



of the TE families (Figure 3); in other words, older families of elements tend to be more modified than younger families. The same analysis was done using the local approach to computing the histone modification background tag counts, as described in the previous section, and the results are qualitatively similar when this technique is applied (Additional file 1, Figure S6). These data are consistent with the exaptation hypothesis for TE modifications, as opposed to the genome defense model, and suggest that many older TE sequences may be preserved, at least in part, due to the contributions they make the epigenetic environment of the human genome.

TE ages and histone modifications

The divergence of an individual TE insertion from its subfamily consensus sequence is a barometer of the time elapsed since its insertion and is, thus, a good measure for its relative age [25]. As shown in Figure 3, a comparison between TE families indicates a positive correlation between element ages and the extent of histone tail modifications. This observation is consistent with the exaptation hypothesis, which predicts that older TEs will bear more epigenetic modifications. However, these results may be confounded by comparisons between families made up of very different kinds of TEs with distinct insertion mechanisms, genomic

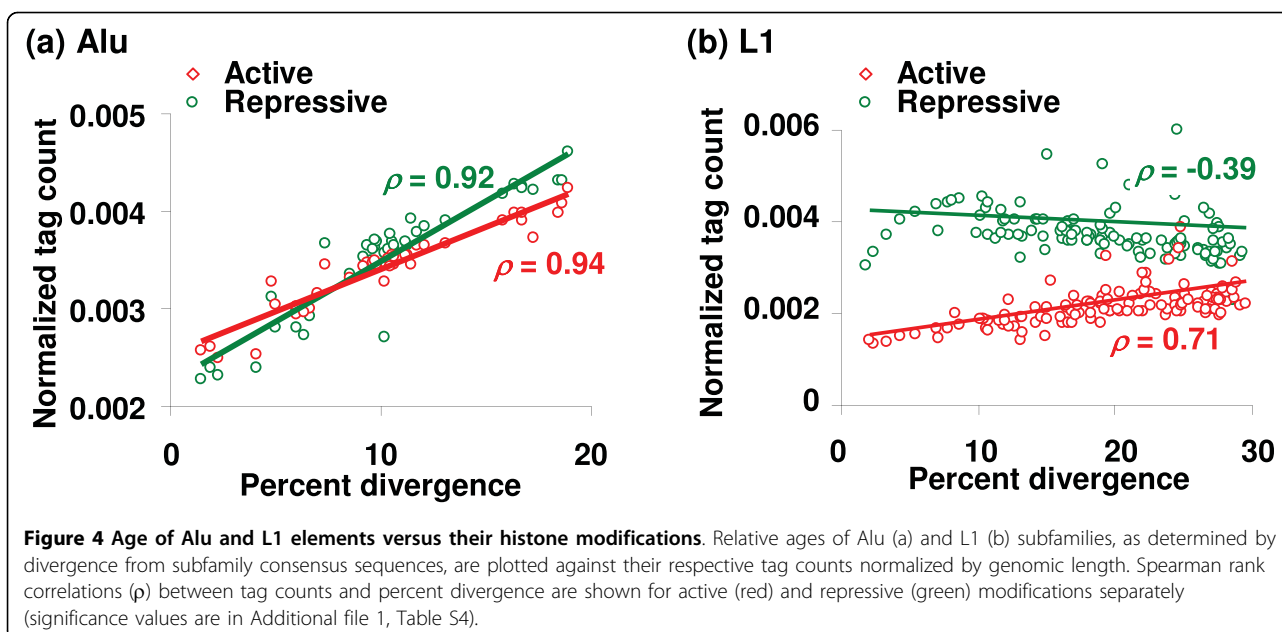
distributions and life histories. In order to evaluate the relationship between element ages and histone tail modifications in a more controlled way, we compared the extent of TE histone modifications with the relative ages of TE insertions within the Alu and L1 families of elements. The Alu and L1 families were chosen for two reasons: first, they are numerous and abundant providing statistical resolution on the question; secondly, and more importantly, they have well-characterized subfamilies the relative ages of which are known [25-27]. The relative ages of individual Alu and L1 insertions can be inferred by comparing their sequences to the consensus sequences of their subfamilies (Additional file 1, Figures S11 and S12) and these data are provided in the output of the RepeatMasker program used to annotate the elements. We computed the average element-to-subfamily consensus sequence divergence for all Alu and L1 subfamilies and compared these values to the extent of active and repressive histone modifications that map to members of the individual subfamilies.

The within-family analyses of the relationship between the relative ages of Alu elements and their histone modifications yield results that are most consistent with the exaptation hypothesis (Figure 4a). Alu element ages are significantly positively correlated with both active ($\rho = 0.94$, $P = 4e-20$) and repressive ($\rho = 0.92$, $P = 9e-18$) histone modifications (Additional file 1, Table S4). These data indicate that members of older Alu subfamilies are subject to more active and repressive modifications, which stands in contrast to the prediction of the genome defense model that younger elements should be more repressed.

The relationships between the ages of L1 elements and their histone modification states appear to support both the genome defense and exaptation models (Figure 4b). The ages of L1 elements are negatively correlated with repressive modifications ($\rho = -0.39$, $P = 5e-6$) and positively correlated with active modifications ($\rho = 0.71$, $P = 4e-20$) (Additional file 1, Table S4). The relative abundance of repressive modifications of younger L1s is consistent with the genome defense model, whereas the data for the increasing active modifications of older L1 elements are consistent with the exaptation model. Taken together, the within-family data for Alu and L1 elements display a complex view of the relationship between TE ages and histone modifications suggesting interplay between the genome defense and exaptation hypotheses.

TE-gene locations and histone modifications

The exaptation hypothesis predicts that TEs proximal to host genes would bear more histone modifications than those that are distal to genes, since these modifications are more likely to effect the regulation of the genes. In order to test this prediction, we analysed the Alu and



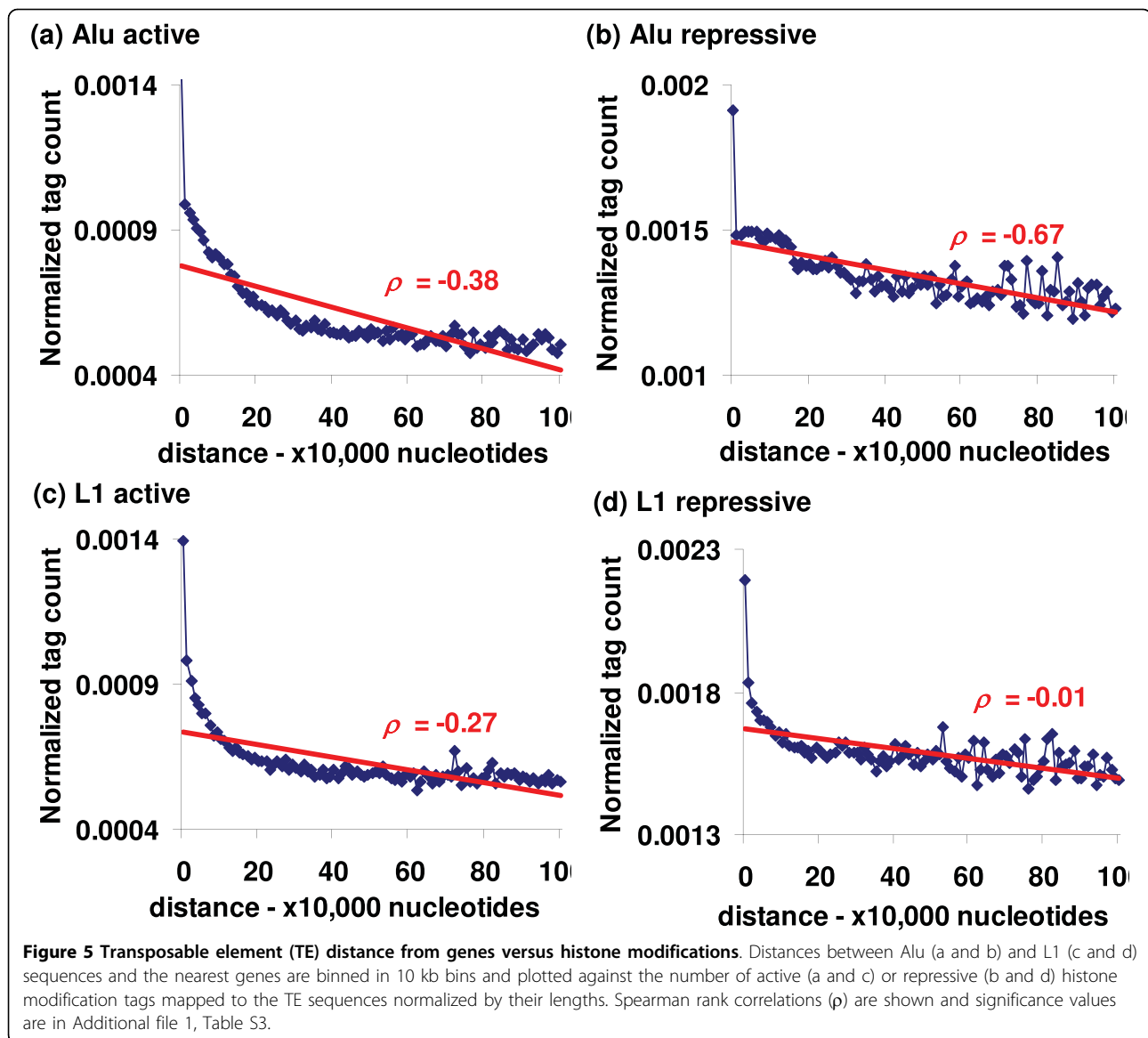
L1 TE families and associated every TE sequence to the nearest gene. The corresponding tag counts of active and repressive histone modifications in TEs were binned according to their distance from genes. Only uniquely mapped TE-tags that could be assigned unambiguous genomic locations were used for this analysis. Alu and L1 were chosen both for their genomic abundance and for the fact that they have distinct genomic distributions: Alus are enriched near genes, whereas L1s are found more often in intergenic regions. For both Alu and L1, we observed negative correlations (Alu active $\rho = -0.38$, $P = 5e-5$, Alu repressive $\rho = -0.67$, $P = 9e-14$, L1 active $\rho = -0.27$, $P = 0.003$, L1 repressive $\rho = -0.01$, $P = 0.46$) between TE insertion distances from genes and histone modifications (Figure 5 and Additional file 1, Table S3). Moreover, TEs that lie within gene boundaries are modified at much higher levels compared to those outside of genes. These findings are in agreement with the exaptation hypothesis. The same analysis was done using both unique and repetitively mapping tags, and the results are qualitatively unchanged when this more comprehensive approach is taken (Additional file 1, Figure S7).

Conclusions

Comparison with previous results

While most work to date on mammalian histone modifications has focused on non-repetitive DNA, there have been four recent studies on the histone modification status of mammalian repetitive sequence elements, three in mouse [8,20,22] and one in human [19]. The previous studies focused on repressive histone modifications and

they turned up a number of cases where mammalian TEs, including SINEs, LTR and DNA elements, were found to be enriched for specific histone modifications. We compare the results of these previous studies with the findings reported here in Table 1. Interestingly, the results reported here agree and disagree with those of previous studies in equal measure. When specific histone modifications are considered for individual TE classes, there are six cases where histone modifications previously identified to be enriched for a given TE class are enriched in the same class in our study, and there are six cases where previously enriched TE-modifications are found to be depleted here. These discrepancies underscore the extent to which histone modifications, particularly those that target TEs, may be cell-type specific, since the different studies that are being compared analysed different cell types. Indeed, the study of Martens *et al.* evaluated multiple cell types and found that histone modifications of TEs were more variable across cell types than those of tandem satellite repeats [20]. This was attributed to the fact that tandemly repeated DNA, such as that found around centromeres, form more stable chromatin architectural elements and tandem repeats are present in more constitutively heterochromatic environments. Interspersed repeats, on the other hand, may be more prone to cell-type specific *in situ* formation of heterochromatic regions dispersed among the euchromatic portion of the genome. This has been seen in plants where insertions of TEs lead to the localized spread of repressive chromatin [2]. In any case, a deeper understanding of how human TEs are epigenetically modified, along with the regulatory implications,



will require a comparison of TE-modifications across a variety of cell types.

Exaptation as a local or global phenomenon

Exaptation refers to the evolutionary process whereby an organismic feature comes to play some role for which it was not originally evolved or selected [5]. TEs are primarily selfish genetic elements that evolved solely virtue of their ability to transpose and thus out-replicate the host genomes in which they reside [28,29]. They do not owe their evolutionary success to any ability to provide functional utility to their hosts. However, at this time it is widely recognized that a number of individual TE sequences have been exapted to play some positive role for their host genomes [6,7]. Exaptation of individual TE sequences may include cases where TEs become incorporated into host protein coding genes or cases where

TEs provide regulatory sequences that help to control the expression of host genes. Such examples of TE exaptation are very much in keeping with the original definition of exaptation as referring to a series of individual, and largely contingent, cases. However, the genome-scale approach taken here to exploring the implications of TE epigenetic modifications entails the consideration of exaptation as a more global, rather than a strictly local, phenomenon. This is because there are particular features of TEs, specifically their ability to recruit epigenetic modifications, which are shared across many elements over the entire genome and which, in turn, allow individual insertions to be exapted. This does not mean that all TEs in the genome are exapted. Rather, the data reported here suggest that there are genome-scale signals, in terms of how the TEs are epigenetically

Table 1 Comparison of transposable element (TE) histone modification enrichments found in this study with those of previous studies.

Enriched in previous study ^a	Status in current study ^b
Kondo and Issa 2003 (Human) [19]	
SINE: H3K9me2	Depleted
Martens <i>et al.</i> 2005 (Mouse) [20]	
SINE: H3K9me3	Depleted
SINE: H3K27me3	Enriched
SINE: H4K20me3	Depleted
LTR: H3K9me3	Enriched
LTR: H3K27me3	Enriched
LTR: H4K20me3	Depleted
DNA: H3K27me3	Enriched
DNA: H4K20me3	Depleted
Mikkelsen <i>et al.</i> 2007 (Mouse) [8]	
LTR: H3K9me3	Enriched
LTR: H4K20me3	Depleted
Pauler <i>et al.</i> 2008 (Mouse) [22]	
SINE: H3K27me3	Enriched

^a TE classes (SINE, LINE, LTR or DNA) that were shown to be enriched for specific histone modifications (as shown) in previous studies.

^b Status of the same TE class-histone modification pairs as enriched or depleted in this study

modified, which indicate an overall potential for individual human TE sequences to be exapted. Consideration of exaptation as a global or genome-scale phenomenon as it relates to TEs reveals how inherent features of the elements, such as their ability to be transcribed or their dispersed repetitive nature, serve to recruit the very epigenetic machinery that will allow them to affect the regulation of nearby genes. Having established this global pattern of TE epigenetic exaptation, further inquiry can now be used to identify individual cases of interest. We give specific examples of how individual cases of TE epigenetic exaptation may be uncovered in the following section.

Caveats and future directions

As mentioned previously, TE epigenetic modifications are certain to be cell-type specific to some extent. Here, we only analysed histone modifications of human TEs in a single cell type - CD4⁺ T cells. As more and more genome-scale histone modification data sets become available, it will become possible to systematically evaluate changes in the histone modification states of TEs across tissues. This is particularly relevant for a deeper interrogation of the genome defense hypothesis. Vertical transmission (inheritance) of novel TE insertions, along with their mutagenic effects, is dependant upon transposition events that occur in the germline, as opposed to TE insertions in somatic tissue, which is an evolutionary dead end. For this reason, one may expect that

the most vigorous genome defense mechanisms would be employed in germline tissue. Thus, it is possible that the predictions of the genome defense model, which are not supported for the most part in this study, may be borne out if germline tissue was evaluated in the same way as done here for somatic tissue. However, there is some evidence that suggests this may not be the case for human TEs. Alu elements, which make up a huge fraction of the methylated DNA in the human genome in somatic tissues, are actually hypomethylated in the male germline [30]. This may represent an evolutionary strategy for the elements, whereby the TEs mitigate their deleterious effects in somatic tissue by reducing transposition therein and yet allow for the transmission of new insertions across generations by relaxing element suppression in the germline [31]. This kind of strategy can be seen for P elements in *Drosophila*, which utilize alternative splicing to encode a repressor protein in somatic tissue and a transposase in the germline [32]. Nevertheless, a better understanding of the role epigenetic histone modifications in the repression of heritable TE insertions will require the analysis of germline tissue.

The genome-wide mapping of 38 histone modifications in the human genome enabled us to thoroughly investigate the relationship between TEs and epigenetic histone modifications. We tested several predictions generated by two competing hypotheses - the genome defense hypothesis and the exaptation hypothesis - in the light of epigenetic histone modifications. Consistent with the exaptation hypothesis, we found that the overall enrichment of histone modifications is positively correlated with the increasing age of TE insertions, and TEs proximal to human genes bear more histone marks than TEs distal to genes. We also found support for the genome defense hypothesis for certain cases, but the majority of our data and analyses support the exaptation hypothesis.

Thus, for the human genome, some epigenetic modifications of TEs may serve to regulate the expression of host genes rather than to silence the elements themselves. More definitive proof of epigenetically related exaptation of TEs will require the analysis of individual cases whereby specific TE sequences have been exapted to regulate host genes. These could include TE-derived promoter sequences, which provide local regulatory sequences and transcription start sites to host genes, and/or TE-derived enhancers that regulate genes from more distal locations. An evaluation of how such TE-derived regulatory sequences are epigenetically modified across different cell types along with an examination of how cell-type specific modifications correspond to expression differences should help to reveal epigenetic routes by which TEs influence their host genomes.

Methods

Tag-to-genome mapping

The genome-wide distributions of 38 histone tail modifications were previously evaluated in human CD4⁺ T cells using ChIP-Seq with the Illumina-Solexa platform [23,24]. The mapping protocol used in these studies did not allow for the consideration of histone modifications at repetitive DNA sequences, since they removed redundantly mapping sequence tags. Therefore, we employed a heuristic mapping procedure for the data generated in these ChIP-Seq studies in order to be able to analyse sequence tags that map to repetitive DNA. To do this, we downloaded 140 sequence tag libraries corresponding to the 38 previously characterized CD4⁺ T cell histone tail modifications from the NCBI Short Read Archive (SRP000200 and SRP000201) [33]. Sequence reads and their respective quality scores were converted from Illumina-Solexa format to the standard (Sanger) fastq format, and the MAQ (Mapping and Alignment with Qualities) program was used to map each fastq library to the March 2006 human genome reference sequence (NCBI Build 36.1, hg18 assembly). MAQ uses a mapping algorithm that utilizes the tag sequences along with their quality scores to determine the highest scoring match to the genomic location [34]. MAQ was run in such a way that tags with more than one identically scoring best tag-to-genome alignment, *i.e.* repetitively mapping tags, were randomly assigned to one genomic location. This procedure allowed us to avoid the elimination of sequence tags that have high scoring tag-to-genome alignments but map to more than one location. Since human TEs can be characterized into related groups (classes, families and subfamilies), using this heuristic mapping procedure provides an unambiguous way to evaluate differences in the frequencies of specific histone modifications between related groups of TEs.

Gene expression-histone modification enrichment analysis

We downloaded the Refseq annotations of experimentally characterized transcription start sites from the database of transcription start sites (DBTSS) [35,36], and mapped them to the human genome reference sequence (hg18) at the UCSC Genome Browser [37]. CD4⁺ T cell expression data corresponding to the mapped Refseq genes were taken from the Novartis Gene Expression Atlas 2 [38]. We were able to obtain unambiguously mapped transcription start sites and gene expression data for 12,644 human genes. We defined promoter regions as 1000 nucleotides upstream and 200 nucleotides downstream of the transcription start sites. We located the number of tags corresponding to each histone tail modifications in each promoter region. The number of tags of each

modification in a promoter region was converted to a binary presence/absence call using a genomic background tag distribution and a conservative threshold determined by the Poisson distribution and incorporating Bonferroni correction for multiple tests [24].

Combining the CD4⁺ T cell gene expression data with promoter histone modification presence/absence calls, we calculated the expression enrichment for each histone modification using the following formula:

$$\text{Expression fold change} = \log_2 \left(\frac{\text{average expression of genes with modification}}{\text{average expression of genes without modification}} \right)$$

In addition, for each histone tail promoter modification, the significance of the difference in average CD4⁺ T cell gene expression levels for genes with and without the modification was evaluated using the Student's *t*-test.

TE-histone modification enrichment analysis

We downloaded RepeatMasker [39] annotations (version 3.2.7) of TE locations for the human genome reference sequence (hg18) from the UCSC genome browser. Using the TE genomic coordinates and our tag-to-genome mapping data, we co-located the tags that correspond to each histone tail modification with TE sequences in the human genome. In this way, we obtained the number of tags of each histone tail modification that map to TE sequences in the human genome.

The TE-histone modification mapping dataset was divided into six classes (families) of TEs [40,41] which are: Alu, MIR, L1, L2, DNA transposons and LTR-retrotransposons. We normalized the number of histone modification tags in each class (family) of TE sequences by the total genomic length of these TE sequences in the class (family), and compared the normalized TE tag counts to either (1) genome-wide background tag counts or (2) locally computed genomic background tag counts. Genome-wide background tag counts are the total number of tags for each modification divided by the length of the genome. To obtain local histone modification background tag counts for TE classes (families), for each individual TE insertion, a background tag count was computed by randomly sampling a non-TE sequence of the same size from within a 1 megabase genomic window surrounding that TE. These individual local background tag counts were then averaged over all TE insertions of a given class (family). The following formulas were used for enrichment calculations:

$$\text{TE fold change}_{\text{Alu, L1, LTR, DNA, MIR, L2}} = \left(\frac{\text{Normalized tag count in TE sequences}}{\text{Normalized tag count in genomic background}} \right)$$

where

$$\text{Normalized tag count}_{\text{modification1-38}} = \frac{\sum \text{tags located in TE sequences}}{\sum \text{length of TE sequences}}$$

Statistical analyses

The statistical significance of TE-histone modification enrichment values were calculated using the goodness of fit *G*-test, which uses a log-likelihood ratio comparing the observed to expected tag counts. The *P*-value thresholds for the *G*-tests were adjusted using the Bonferroni correction for multiple tests. Prior to correlation analysis, all data distributions were checked for normality using Q-Q plots to visually compare the observed distributions against theoretical normal distributions (Additional file 1, Figures S8-S10). Data with distributions that were deemed to be normal were correlated using Pearson correlation (*r*) and data with distributions that were deemed to be non-normal were correlated using Spearman rank correlation (*ρ*). Note that when data are binned, such as for the distance from gene computation, correlations are calculated on the unbinned data. Statistical significance values for correlations were computed using an approximation to the Student's *t*-distribution with *n*-2 degrees of freedom [42].

Additional file 1: Supplementary material. Figures S1-12 and Tables S1-4 are included in the supplementary material file. Click here for file
[http://www.biomedcentral.com/content/supplementary/1759-8753-1-2-S1.PPT]

Abbreviations

ChIP-Seq: chromatin immunoprecipitation followed by high-throughput sequencing; ETn: early transposon; GC: guanine-cytosine; IAP: intracisternal A particle; LTR: long terminal repeat; MIR: mammalian-wide interspersed repeat; SINE: short interspersed nuclear element; TE: transposable elements.

Acknowledgements

This research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI. LMR is supported by Corporacion Colombiana de Investigacion Agropecuaria - CORPOICA. IKJ and graduate student AH were supported by an Alfred P Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). AH was supported by the School of Biology at the Georgia Institute of Technology. The authors would like to thank Lee S Katz and Troy Hilley for helpful discussions and technical advice. The authors would also like to thank Keji Zhao and Chongzhi Zang for providing assistance with the procurement of their dataset.

Author details

¹School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30332, USA. ²National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. ³Computational Biology and Bioinformatics Unit, Biotechnology and Bioindustry Center, Corporacion Colombiana de Investigacion, Agropecuaria - CORPOICA, Km 14 Via a Mosquera, Bogota, Colombia.

Authors' contributions

IKJ and AH conceived of and designed the study, performed computational analyses and wrote up the results. LMR provided technical expertise and assistance for dataset acquisition, curation and analysis. All authors read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 19 June 2009

Accepted: 25 January 2010 Published: 25 January 2010

References

- Matzke MA, Mette MF, Matzke AJ: Transgene silencing by the host genome defense: implications for the evolution of epigenetic control mechanisms in plants and vertebrates. *Plant Mol Biol* 2000, **43**:401-415.
- Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R: Role of transposable elements in heterochromatin and epigenetic control. *Nature* 2004, **430**:471-476.
- McDonald JF, Matzke MA, Matzke AJ: Host defenses to transposable elements and the evolution of genomic imprinting. *Cytogenet Genome Res* 2005, **110**:242-249.
- Yoder JA, Walsh CP, Bestor TH: Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* 1997, **13**:335-340.
- Gould SJ, Vrba ES: Exaptation: a missing term in the science of form. *Paleobiology* 1982, **8**:4-15.
- Kidwell MG, Lisch DR: Transposable elements and host genome evolution. *Trends Ecol Evol* 2000, **15**:95-99.
- Feschotte C: Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 2008, **9**:397-405.
- Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007, **448**:553-560.
- Silva JC, Shabalina SA, Harris DG, Spouge JL, Kondrashov AS: Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet Res* 2003, **82**:1-18.
- Gendrel AV, Lippman Z, Yordan C, Colot V, Martienssen RA: Dependence of heterochromatic histone H3 methylation patterns on the Arabidopsis gene DDM1. *Science* 2002, **297**:1871-1873.
- Grewal SI, Elgin SC: Transcription and RNA interference in the formation of heterochromatin. *Nature* 2007, **447**:399-406.
- Grewal SI, Jia S: Heterochromatin revisited. *Nat Rev Genet* 2007, **8**:35-46.
- Lippman Z, Martienssen R: The role of RNA interference in heterochromatic silencing. *Nature* 2004, **431**:364-370.
- Slotkin RK, Martienssen R: Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* 2007, **8**:272-285.
- Suzuki MM, Bird A: DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* 2008, **9**:465-476.
- Weil C, Martienssen R: Epigenetic interactions between transposons and genes: lessons from plants. *Curr Opin Genet Dev* 2008, **18**:188-192.
- Zaratiegui M, Irvine DV, Martienssen RA: Noncoding RNAs and gene silencing. *Cell* 2007, **128**:763-776.
- Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI, Martienssen RA: Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* 2002, **297**:1833-1837.
- Kondo Y, Issa JP: Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells. *J Biol Chem* 2003, **278**:27658-27662.
- Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, Steinlein P, Jenuwein T: The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *Embo J* 2005, **24**:800-812.
- Bernstein BE, Meissner A, Lander ES: The mammalian epigenome. *Cell* 2007, **128**:669-681.
- Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, Sommer A, Aszodi A, Jenuwein T, Barlow DP: H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res* 2009, **19**:221-33.

23. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129**:823-837.
24. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K: **Combinatorial patterns of histone acetylations and methylations in the human genome.** *Nat Genet* 2008, **40**:897-903.
25. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
26. Jurka J: **Subfamily structure and evolution of the human L1 family of repetitive sequences.** *J Mol Evol* 1989, **29**:496-503.
27. Kapitonov V, Jurka J: **The age of Alu subfamilies.** *J Mol Evol* 1996, **42**:59-65.
28. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution.** *Nature* 1980, **284**:601-603.
29. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite.** *Nature* 1980, **284**:604-607.
30. Chesnokov IN, Schmid CW: **Specific Alu binding protein from human sperm chromatin prevents DNA methylation.** *J Biol Chem* 1995, **270**:18539-18542.
31. Bowen NJ, Jordan IK: **Transposable elements and the evolution of eukaryotic complexity.** *Curr Issues Mol Biol* 2002, **4**:65-76.
32. Rio DC: **Molecular mechanisms regulating Drosophila P element transposition.** *Annu Rev Genet* 1990, **24**:543-578.
33. NCBI Short Read Archive. <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?>
34. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18**:1851-1858.
35. Huda A, Marino-Ramirez L, Landsman D, Jordan IK: **Repetitive DNA, nucleosome binding and human gene expression.** *Gene* 2009, **436**:12-22.
36. Suzuki Y, Yamashita R, Nakai K, Sugano S: **DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs.** *Nucleic Acids Res* 2002, **30**:328-331.
37. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31**:51-54.
38. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101**:6062-6067.
39. RepeatMasker. <http://www.repeatmasker.org>.
40. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110**:462-467.
41. Kapitonov VV, Jurka J: **A universal classification of eukaryotic transposable elements implemented in Repbase.** *Nat Rev Genet* 2008, **9**:411-412.
42. Sokal RR, Rohlf JF: *Biometry: The Principles and Practice of Statistics in Biological Research* San Francisco: W. H. Freeman 1981.

doi:10.1186/1759-8753-1-2

Cite this article as: Huda *et al.*: Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mobile DNA* 2010 **1**:2.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp



Many sequence-specific chromatin modifying protein-binding motifs show strong positional preferences for potential regulatory regions in the *Saccharomyces cerevisiae* genome

Loren Hansen^{1,2}, Leonardo Mariño-Ramírez³ and David Landsman^{1,*}

¹Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8900 Rockville Pike, Bethesda, MD 20894, ²Bioinformatics Program, Boston University, Boston, MA 02215, USA and ³Computational Biology and Bioinformatics Unit, Plant Molecular Genetics Laboratory, Biotechnology and Bioindustry Center, Corporacion Colombiana de Investigacion Agropecuaria – CORPOICA Bogota, Colombia

Received October 9, 2009; Revised December 7, 2009; Accepted December 8, 2009

ABSTRACT

Initiation and regulation of gene expression is critically dependent on the binding of transcriptional regulators, which is often temporal and position specific. Many transcriptional regulators recognize and bind specific DNA motifs. The length and degeneracy of these motifs results in their frequent occurrence within the genome, with only a small subset serving as actual binding sites. By occupying potential binding sites, nucleosome placement can specify which sequence motif is available for DNA-binding regulatory factors. Therefore, the specification of nucleosome placement to allow access to transcriptional regulators whenever and wherever required is critical. We show that many DNA-binding motifs in *Saccharomyces cerevisiae* show a strong positional preference to occur only in potential regulatory regions. Furthermore, using gene ontology enrichment tools, we demonstrate that proteins with binding motifs that show the strongest positional preference also have a tendency to have chromatin-modifying properties and functions. This suggests that some DNA-binding proteins may depend on the distribution of their binding motifs across the genome to assist in the determination of specificity. Since many of these DNA-binding proteins have chromatin remodeling properties, they can alter the local nucleosome structure to a more permissive and/or restrictive state, thereby assisting in determining DNA-binding protein specificity.

INTRODUCTION

At any given point in time, cells are performing complex programs of gene expression. The binding of transcriptional regulators to target genes determines their expression or repression. Many DNA-binding proteins (DBPs) recognize and bind specific DNA sequence motifs located within specific regulatory regions of the gene. However, the length and nucleic acid composition of these binding motifs frequently enables their random occurrence within the genome, sometimes up to thousands of repetitions. Therefore, sequence information alone is insufficient to completely determine specificity (1,2).

Within the nucleus, DNA exists in complexes with RNA and proteins called chromatin. Commonly composed of an octamer of histone proteins consisting of two copies each of histones H2A, H2B, H3 and H4, nucleosomes are the basic repeating units of chromatin [for review see ref. (3)]. DNA wraps around the histone octamer core in approximately two superhelical turns. These cores are spaced ~10–80 bp apart; this internucleosomal DNA is referred to as linker DNA. This DNA can vary in length significantly, even between neighboring nucleosomes. DNA within nucleosomes is less accessible to DBPs, including transcriptional regulators (4). It has long been thought that by occupying potential binding sites, nucleosomes play an indirect role in regulating gene expression (4–7). However, this raises the question of how the structure of chromatin is constructed initially to ensure the availability of sites for transcriptional regulator binding. It is likely that inherent signals within the DNA sequence play an important role in positioning nucleosomes (8,9). Also critical are chromatin remodeling factors (CRFs) that reposition or modify nucleosomes (8,10–13), thereby repressing or

*To whom correspondence should be addressed. Tel: +1 301 435 5981; Fax: +1 301 480 2288; Email: landsman@ncbi.nlm.nih.gov

enhancing transcription. Whether and how CRFs act to modify chromatin structure to a more permissive/restrictive state remains unknown. One possibility is that CRFs rely on the quality and genomic position of their DNA sequence motifs to help establish specificity. In this study, we investigated this hypothesis by examining the positional distribution of predicted binding sites for 184 DBPs in the *Saccharomyces cerevisiae* genome.

MATERIALS AND METHODS

Calculating promoter enrichment scores

Transcription start sites (TSSs), as well as promoter and coding sequences, were obtained from the UCSC genome browser (14). The Mining Yeast Binding Sites (MYBS) database was used to obtain 666 position weight matrices (PWMs) (15). The Spt10 PWM was obtained from ref. (16) for a total of 667 PWMs. Promoters were defined as regions extending 1000 bp upstream of TSSs, excluding any coding sequence. Each PWM was used to score both promoter and coding sequences while looking for subsequences that closely match the binding motif represented by the PWM. The score of each subsequence was derived from the sum of the position-specific score of each nucleotide composing the subsequence. For a subsequence of length $l(s_1 \dots s_l)$ with length l equal to the number of columns in the PWM, the score was calculated as

$$\text{Score} = \sum_{j=1}^l m_{s_j, j} \quad 1$$

where S_j represents the nucleotide at position j of subsequence s and $m_{i,j}$ represents the score in the PWM for row i and column j .

We randomized the sequence of interest by shuffling the nucleotides while retaining the overall nucleotide composition. Then each set of randomized sequence was scanned against the set of PWMs and the number of high-scoring matches was counted. The randomization was performed 800 times, and the mean and standard deviation for the number of matches expected in the randomized sequence for a given PWM was calculated. A z -score representing the degree of sequence motif enrichment was calculated using

$$Z = \frac{x - u_r}{\sigma_r}, \quad 2$$

where x is the number of high-scoring matches for the unshuffled sequence, u_r is the mean number of high-scoring matches for 800 sets of shuffled sequences, and σ_r is the standard deviation for the group of 800 sets of shuffled sequences. We then used the calculated z -scores from the promoter and coding sequence to calculate a promoter enrichment score (i.e. promoter z -score – ORF z -score) for each PWM.

To perform this analysis, it was necessary to select a cutoff score. Therefore, similar to other comparable studies (17), a cutoff score representing 70% of the maximum possible score for a given PWM was chosen. Results from analyses using cutoff scores representing 80

and 90% of the maximum possible score showed little differences.

Gene ontology (GO) analysis

The set of PWMs was filtered using the methods outlined below and then ranked according to the promoter preference score. Finally, using the online David GO tool, we searched for enriched GO terms (18) in the top 20% of PWMs ($N = 37$). As a control, we assessed the set of all proteins (184) represented by the collection of 667 PWMs used in this study. To avoid the use of an arbitrary percentile cutoff, we also applied the online Gene Ontology enrichment anaLysis and visualizaTion (GORilla) tool (19) to our set of ranked proteins. GORilla uses a flexible threshold technique to search for GO terms enriched in a ranked list.

The set of PWMs used contained considerable redundancy (i.e. many DBPs are associated with multiple PWMs). To perform the GO analysis, it was necessary to filter the set of 667 PWMs to obtain a unique set of 184 PWMs to pair with the 184 unique proteins. Two different filtering methods were used to determine which PWM out of the set of PWMs associated with a given DBP would be used when ranking the protein. With the first method, we filtered PWMs based on the promoter enrichment score. The PWM with the highest promoter enrichment score from the set of PWMs was selected to pair with that protein. Each protein was then ranked according to the promoter enrichment score of its paired PWM and GO analysis performed as outlined above. Using this method, both analysis tools identified GO terms related to chromatin modification for the highly ranked proteins. With the second method, we filtered the PWMs according to information content. The PWM with the highest information content was selected to pair with its associated protein. We repeated the above analysis using both GO tools. Using the David tool, we again identified an enrichment of chromatin modifying GO terms for highly ranked proteins ($P < 0.05$). However, GORilla did not reveal any GO terms possibly due to the stringent cutoff ($P < 0.001$) of this tool.

Nucleosome overlap score

With the set of high-scoring matches in promoter regions and a map of nucleosome positions produced in a recent study (20), we calculated the fraction of predicted binding sites that overlapped with a well-positioned nucleosome for each PWM. Nucleosomes, unlike many DBPs, do not necessarily have a well-defined binding site. Instead, they may have multiple binding locations in different cells for the same nucleosome. For each nucleosome, Mavrich *et al.* (20) calculated a 'fuzziness score' that represented the extent a nucleosome varies its binding location. To obtain a list of well-positioned nucleosomes we ranked all nucleosomes by their fuzziness score and took the top 15%.

To calculate the significance of the observed overlap of predicted binding sites with well-positioned nucleosomes, we randomly changed the positions of the predicted binding sites within a 1000-bp window and calculated

the fraction of randomized sites that overlapped with a well-positioned nucleosome. After 1000 iterations, the mean and standard deviation of nucleosome overlap were estimated. In addition, a concurrent z -score representing the degree of nucleosome overlap above or below random chance was calculated according to Equation (2), where x was the fraction of high-scoring matches that overlapped a nucleosome, u_r was the mean fraction of high-scoring matches that overlap a nucleosome calculated based on 1000 random permutations, and σ_r represented the standard deviation of the fractional overlap of the randomly moved high-scoring matches.

Promoter regions have a tendency to contain nucleosome-depleted regions (21). To control for potential bias, we randomly changed the predicted binding site location within a 1000-bp window that was centered on the binding site. In doing so, the randomly permuted binding sites were still mostly positioned within the same local chromatin structure. A 1000-bp window will almost always include some of the neighboring ORF sequences. Thus, while restricting the randomization to a defined window reduces the effect of simply being within a promoter region, it does not eliminate it entirely. One could argue that our results indicating a strong bias toward promoter regions for some motifs exacerbate this issue. However, in our calculation of nucleosome occupancy, we only used those sites found in promoter regions. Hence, promoter bias should not play a significant role in these analyses. For each PWM we paired its promoter enrichment score with its nucleosome overlap score and calculated the correlation using Spearman rank correlation. Correlation coefficients were calculated using those PWMs with at least 50 predicted binding sites.

Within promoter positional analysis

For each high-scoring promoter region match, we calculated the distance to the closest TSS. Predicted binding sites that could not be mapped to a TSS were discarded. Sequence motifs that were highly 'location constrained' within promoter regions clustered together. For every PWM that had at least 50 predicted binding sites within promoter regions, we obtained the distance from the TSS for every high-scoring match (i.e. predicted binding site) and then calculated the mean, median and semi-interquartile range for the distance distribution. The smaller the semi-interquartile range, the more clustered the predicted binding sites were and the stronger the location constraint within promoter regions.

RESULTS

Many DBP sequence motifs displayed strong preferences for promoter regions as opposed to coding regions

Sequence motifs for DBPs are commonly represented by a position weight matrix (PWM) (1,22). We obtained a set of 667 PWMs representing binding motifs for 184 DBPs from the MYBS database (15). For each PWM we calculated a promoter enrichment score. The larger the

score, the more enriched the sequence motif was in promoter regions relative to coding regions.

Not surprisingly, most sequence motifs showed dramatically greater enrichment in promoter regions than in coding regions (Figure 1). For example, Orc1p, which has been demonstrated to function in chromatin modification (23), displayed the greatest difference in enrichment between promoter and coding sequence. For this sequence motif, the number of high-scoring matches within the promoter region was 1240, corresponding to a z -score of 261. Meanwhile, the number of high-scoring sequence motif matches within coding sequence was 38, corresponding to a z -score of -0.88 . Yeast contains ~ 8.4 Mb of coding sequence compared to ~ 2.5 Mb of promoter sequence. Despite this, the Orc1p motif occurred far more often in potential regulatory, but not coding, sequence in the yeast genome.

Sequence motifs showing a strong positional preference were also enriched for CRFs

We then investigated whether proteins whose sequence motifs showed a high positional preference for promoter regions also shared common biological functions. To explore this question, the set of 184 proteins was sorted according to the promoter enrichment score from largest to smallest ('Materials and methods' section). Then the online David bioinformatics resource tool (<http://david.abcc.ncifcrf.gov/home.jsp>) (18) was used to assess GO terms associated with the top 20% of ranked proteins. Chromatin remodeling-related terms were highly represented among these highly ranked proteins ($P < 0.05$), including chromatin modification, establishment and/or maintenance of chromatin architecture, DNA packaging, gene silencing, negative regulation of gene expression epigenetic, chromatin silencing and heterochromatin formation.

To verify these results, we performed a similar analysis using the GOrilla tool (<http://cbl-gorilla.cs.technion.ac.il/>) (19). When given a ranked list of genes, GOrilla searches for GO terms that show greater enrichment for items near the top of the list relative to the rest of the list. Therefore, it was unnecessary to limit this analysis to the top 20% of ranked proteins. We submitted to GOrilla a set of proteins ranked according to their promoter enrichment score and examined GO term enrichment. Similar to the analysis using David, many chromatin-associated GO terms were identified for high-ranking proteins, including histone modification, covalent chromatin modification, and chromatin modification. This analysis indicates that DBPs whose sequence motifs showed the strongest positional constraint for promoters were also associated with CRFs.

A negative correlation exists between high positional preference and nucleosome occupancy

The relationship revealed above between the positional preference of sequence motifs and CRFs led us to postulate that a correlation may also exist between the binding of proteins exhibiting a high positional preference and nucleosome occupancy. Based on nucleosome positions

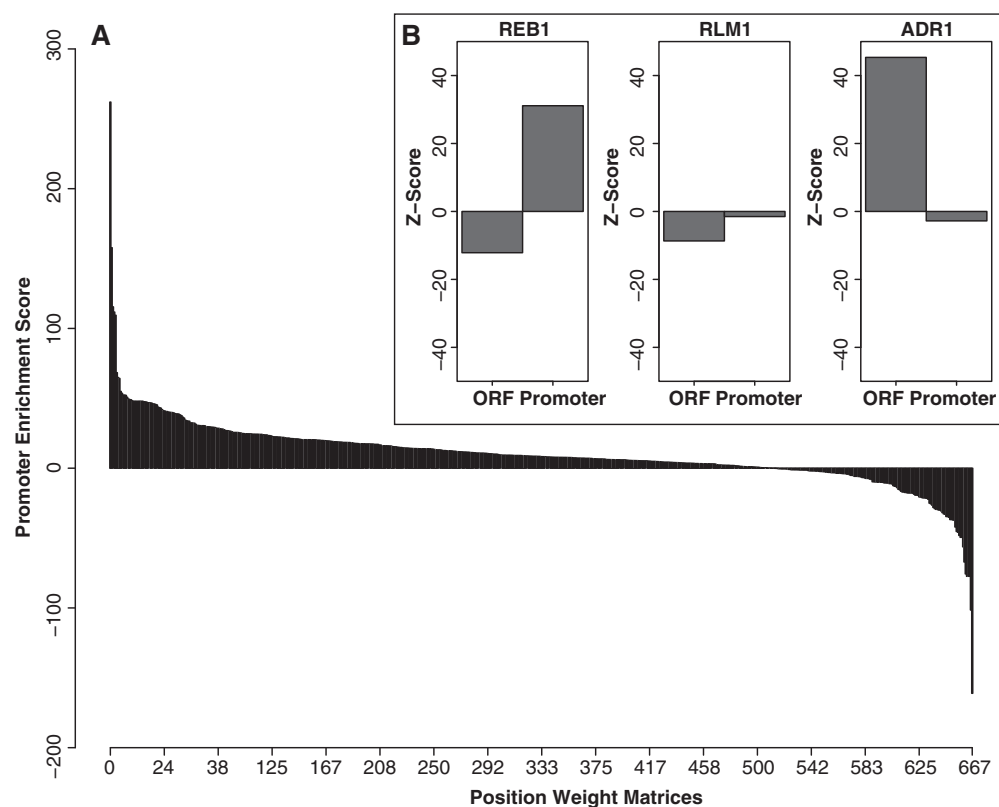


Figure 1. Promoter enrichment scores for all 667 PWMs. (A) The promoter enrichment score for all 667 PWMs sorted by decreasing promoter enrichment. (B) Examples of specific PWMs taken from the top, middle and end of the set of PWMs sorted by promoter enrichment. From left to right, the first graph is a random example from the top 50 PWMs (REB1). The second graph (RLM1) is a random example selected from PWMs ranked 300–400. The third graph (ADR1) is a random example taken from PWMs ranked 600–667. Plotted in (B) is the z-score for each PWM indicating over- or under-representation in the given class of sequence elements. A positive z-score denotes over-representation while a negative z-score signifies under-representation.

obtained in a recent Chip-Seq study (20), we calculated a score to represent nucleosome occupancy (see ‘Materials and methods’ section) for each PWM.

A large negative score indicated that the overlap between predicted binding sites and nucleosomes was much less than would be expected by random chance. Conversely, a large positive score suggested that the likelihood of an overlap was greater than random chance. The Spearman rank correlation coefficient between the promoter enrichment score and the score representing nucleosome occupancy of predicted binding sites was then calculated. Indeed, there was a negative correlation between positional preference and nucleosome occupancy ($r_s = -0.39$, $P < 1e-16$) (Figure 2A). The P -values for correlation coefficients were calculated according to Best and Roberts (24). This result, combined with those from the GO analysis, suggests that DBPs whose binding sites show strong positional preference may act in part to remove or shift nucleosomes upon binding to allow entry by other transcriptional regulators (10), thereby playing a role in determining specificity.

To further confirm these results, we repeated the correlation analysis using a different measure of nucleosome occupancy. Kaplan *et al.* (8) produced a high-resolution map of nucleosome occupancy across the yeast genome. For each position in the genome, a nucleosome occupancy

score was calculated. A negative number indicated that nucleosome occupancy was below the genome average, while a positive number represented an above average likelihood for occupancy. We obtained the data set from Kaplan *et al.* (8) and averaged the nucleosome occupancy score for the set of predicted binding sites in promoter regions for a given PWM. Then, the Spearman rank correlation between the promoter enrichment score and the average nucleosome occupancy was calculated. With this method, we again observed a correlation between nucleosome occupancy and promoter preference ($r_s = -0.44$, $P < 1e-16$) (Figure 2B).

Kaplan *et al.* also produced a map of nucleosome occupancy for chromatin that was reconstituted *in vitro*. Our results suggest that the trend toward lower nucleosome occupancy for motifs with a high positional preference may be due to active chromatin remodeling by the transcription factors that bind those motifs. As such, we would expect to observe a positive correlation between positional preference and those motifs that showed the largest difference between *in vitro* and *in vivo* nucleosome occupancy. To test this hypothesis, we calculated the correlation between the promoter enrichment score and the difference in nucleosome occupancy *in vitro* and *in vivo* for the set of predicted binding sites in promoter regions for each PWM. As anticipated, promoter enrichment and the

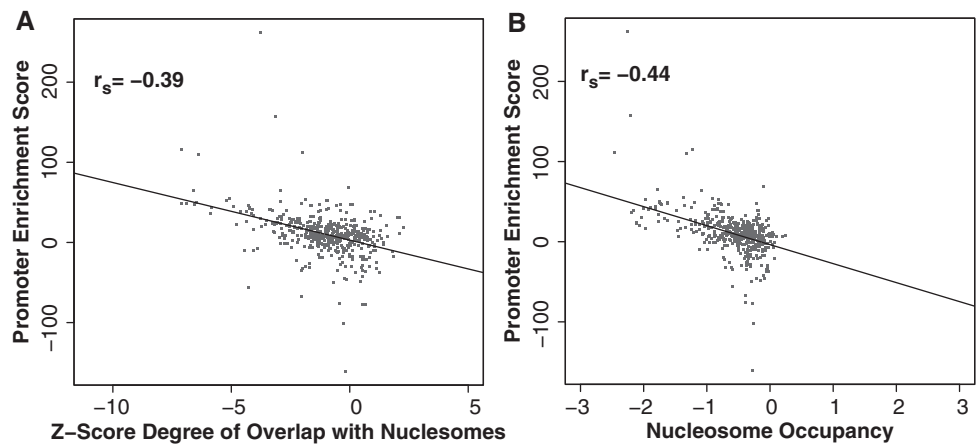


Figure 2. Scatter plots showing the correlation between promoter enrichment and nucleosome occupancy for predicted sites in promoter regions. The y-axis represents promoter enrichment (the larger the value, the stronger the positional preference for promoter regions). The x-axis is a score representing the degree of nucleosome occupancy. For each PWM with at least 50 predicted binding sites in promoter regions, the promoter enrichment score was plotted against the degree of nucleosome occupancy for predicted binding sites using that PWM. (A) A scatter plot using the overlap of predicted transcription factor binding sites with well-positioned nucleosomes as the measure of nucleosome occupancy. (B) The method for measuring nucleosome occupancy as described by Kaplan *et al.* (8).

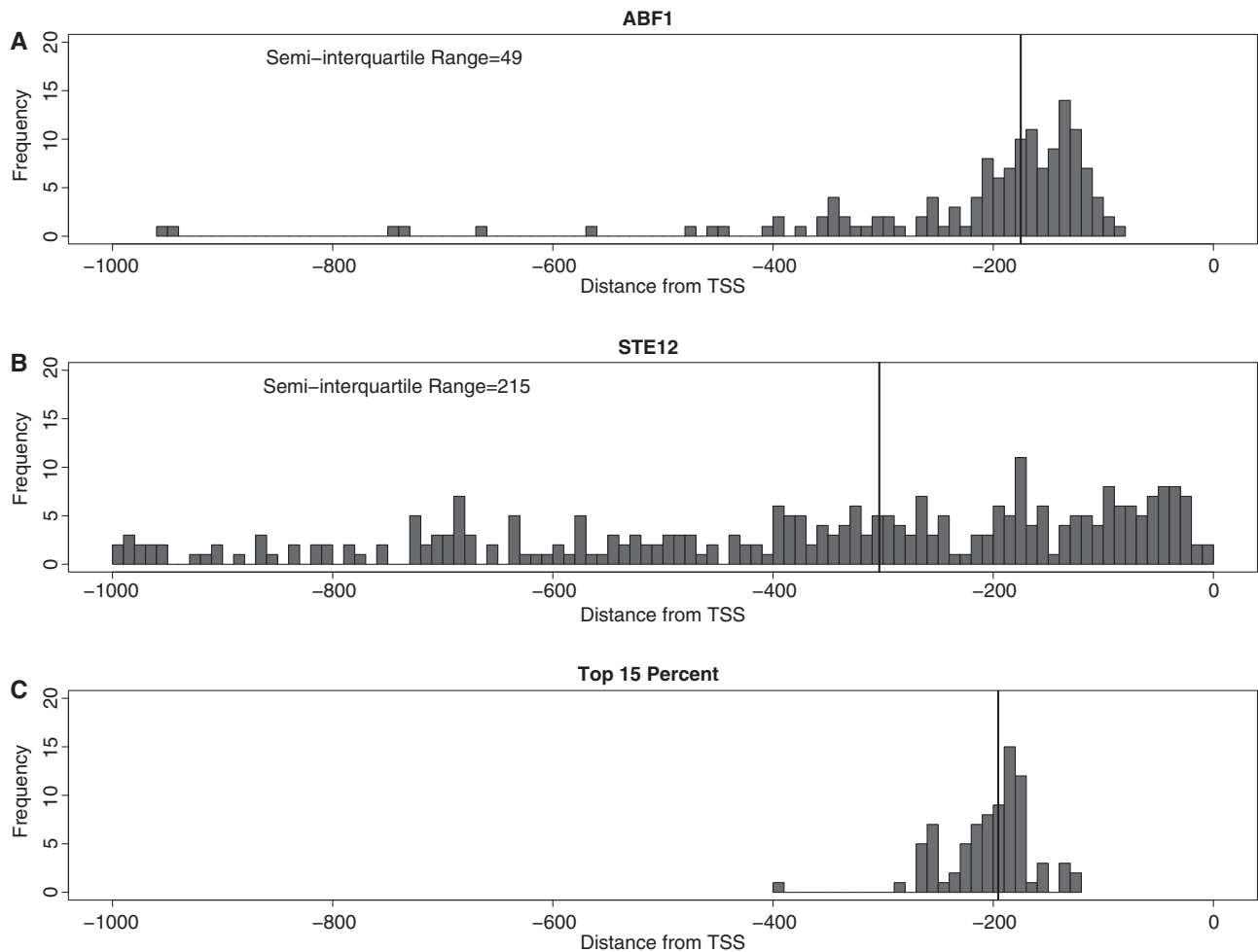


Figure 3. Distribution of high-scoring matches in promoter regions. (A) Example of a sequence motif (ABF1) that is location constrained within promoter regions. Plotted are histograms of distances from the TSS for all predicted binding sites for the indicated DNA-binding protein. The TSS is marked by position zero and the black line represents the median (x-axis units are in bps). (B) Example of a sequence motif that is not location constrained within promoter regions (STE12). (C) All sequence motifs that had at least 50 predicted binding sites were ranked on the basis of semi-interquartile range. Plotted are the medians of the distance distributions for the top 15% ($N = 83$), representing the sequence motifs with a strong promoter positional bias.

difference between *in vitro* and *in vivo* nucleosome occupancy was positively correlated ($r_s = 0.46$, $P < 1e-16$, see Supplementary Figure 1).

A correlation exists between high promoter enrichment and strong location constraint within promoters

Previous studies have shown that motif context, including distance from the TSS, likely plays a role in gene regulation in yeast and humans (25,26). This prompted us to

investigate whether sequence motifs showing strong promoter enrichment also display a strong positional constraint within promoter regions. To answer this question, we calculated the distance to the TSS for predicted binding sites in yeast promoters. Sequence motifs that demonstrated significant location constraint within promoter regions clustered together at similar distances from the TSS corresponding to a narrow distribution of distances (Figure 3A). Sequence motifs that were not constrained within the promoter exhibited distance distributions with

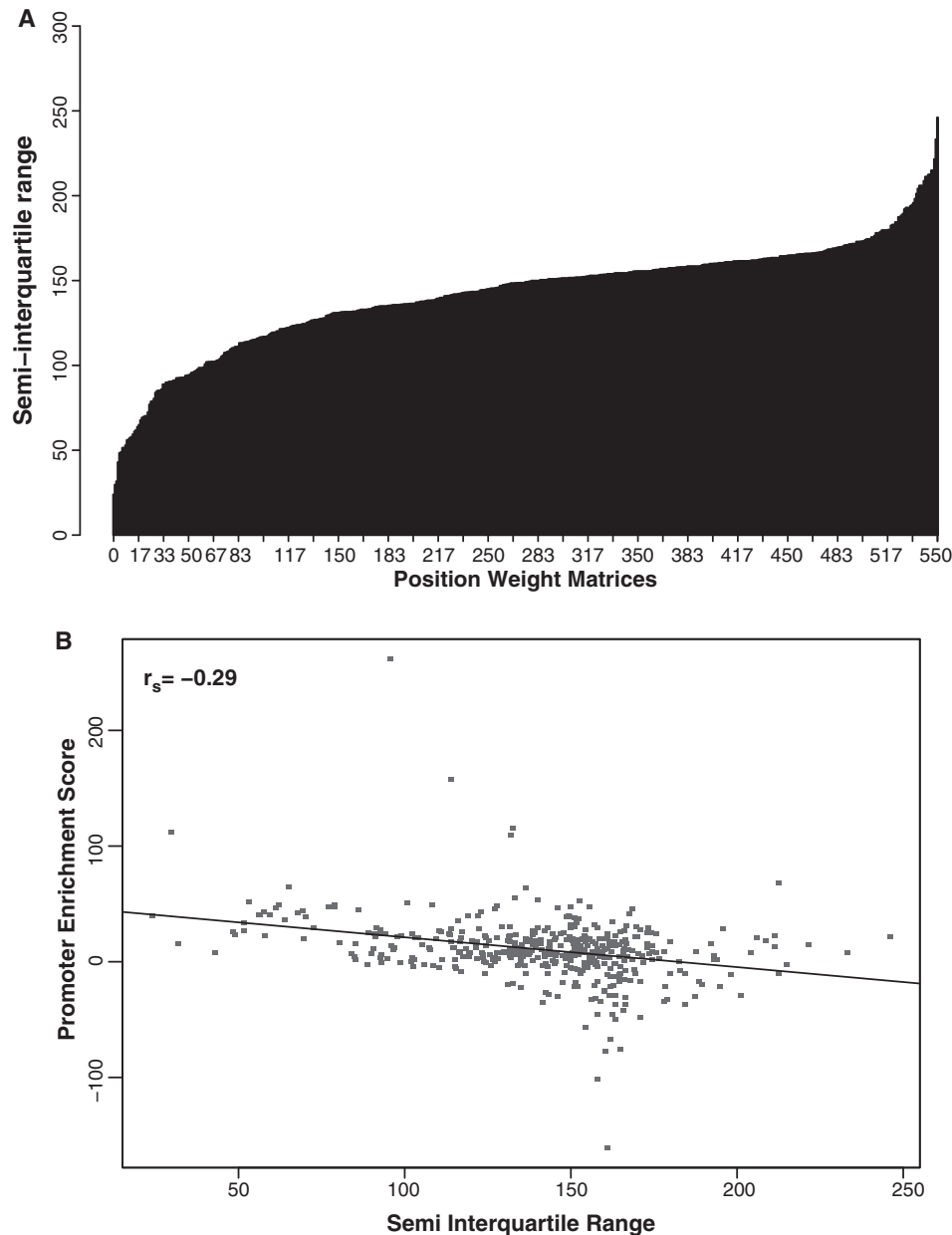


Figure 4. Plot of positional preference within promoter regions. **(A)** The semi-interquartile range for the set of PWMs with at least 50 predicted binding sites within promoter regions ($N = 551$) sorted by increasing semi-interquartile range. The semi-interquartile range measured the distribution dispersion. The larger the value, the greater the distribution spread. A smaller semi-interquartile range indicates more location constraint within promoter regions for the predicted binding sites. **(B)** A scatter plot depicting the correlation between promoter enrichment and positional preference within promoter regions. The y-axis represents promoter enrichment in which larger values signify greater enrichment in promoter regions relative to coding regions. The x-axis represents the degree of positional preference within promoter regions. The smaller the value the more clustered the predicted binding sites are within promoter regions, and the higher the degree of positional preference within promoter regions.

a larger spread (Figure 3B). We noticed with interest that sequence motifs with a strong positional bias within promoter regions seem to cluster ~100–300 bp upstream of the TSS (Figure 3C).

The semi-interquartile range was calculated to measure the distribution spread statistically. Because many of the distance distributions were skewed (see Figure 3a), the semi-interquartile range was a better measure of spread than standard deviation. The Spearman rank correlation coefficient between the positional preference score and the semi-interquartile range was calculated. Indeed, a correlation between positional preference for promoter regions (high promoter enrichment) and positional preference within promoter regions ($r_s = -0.29$, $P = 2.4e-12$) (Figure 4) was revealed.

DISCUSSION

Recent work elucidating nucleosome positioning in yeast has revealed a common chromatin architecture around TSS's consisting of a nucleosome covering the TSS, an immediate upstream nucleosome-free region (NFR) of ~140 bp, and a well-positioned nucleosome (‘-1’ nucleosome) on the upstream border of the NFR (7,27). Veners *et al.* (28) demonstrated that the -1 nucleosome is evicted upon recruitment of RNA polymerase II. Additionally they showed that a number of chromatin remodeling complexes were selectively associated with the -1 nucleosome. Furthermore, a number of sequence-specific experimentally determined binding sites overlapped the -1 nucleosome. These results support the idea that the positioning of the -1 nucleosome may be strongly regulated.

Here we show that sequence motifs with a strong positional bias within promoter regions cluster almost exclusively ~100–300 bp upstream of the TSS (Figure 3C). This localization places them in a prime location to regulate or be regulated by the -1 nucleosome, further supporting the idea that positioning of the -1 nucleosome is important in transcriptional regulation.

If CRFs with sequence motifs that exhibit strong positional preferences are modifying the chromatin structure in part to provide specificity to other DBPs, what is the mechanism of action? One possibility is that CRFs remove and/or shift nucleosomes to open up binding sites for other transcriptional regulators. For example, Rap1p, Abf1p and Reb1p are all highly abundant sequence-specific general regulatory factors that bind motifs with a strong preference for promoter regions. There is good evidence that all three play a role in influencing chromatin structure (10,29,30). Additionally, these proteins appear to act in part by creating bubbles of open chromatin (8,31–33). In the case of Rap1p and Abf1p, creating a region of open chromatin appears to facilitate the binding of additional regulatory factors, leading to transcription enhancement (31). In many cases, Rap1p and Abf1p are unable to activate robust transcription alone (34,35) and require additional regulatory factors. Further support is provided by the observation that Rap1p- and Abf1p-binding sites can be

substituted for one another without a loss in function (31,35).

However, both Rap1p and Abf1p are involved in many functions, including repression (36–38). Rap1p initiates a repressive chromatin structure by interacting directly with the chromatin modifying factors Sir3p and Sir4p (37). Therefore, in addition to making binding sites accessible, it is likely that DBPs whose sequence motifs show a strong positional preference can increase specificity by directly interacting with chromatin modifiers or transcriptional regulators.

A question that immediately presents itself is whether or not the pronounced preference for promoter regions is sufficient to determine specificity. Is the positional distribution sufficient to fully explain binding *in vivo*? In a genome-wide location analysis, Lieb *et al.* (39) noted the strongly skewed positional preference of Rap1p-binding motifs and concluded that the positional distribution of potential Rap1p-binding sites may account for much of the specificity in Rap1p binding. However, the skewed positional distribution of these potential binding sites was insufficient in fully explaining the pattern of Rap1p binding. For the case of Rap1p, additional genome-wide mechanisms also appear to be at work.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank John Spouge for reviewing the manuscript and for his suggestions.

FUNDING

L.M.R. was supported by CORPOICA. Funding for open access charge: Intramural Program of the National Library of Medicine, National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. D'Haeseleer, P. (2006) What are DNA sequence motifs? *Nat. Biotechnol.*, **24**, 423–425.
2. Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
3. Felsenfeld, G. and Groudine, M. (2003) Controlling the double helix. *Nature*, **421**, 448–45.
4. Liu, X., Lee, C.K., Granek, J.A., Clarke, N.D. and Lieb, J.D. (2006) Whole-genome comparison of Leu3 binding *in vitro* and *in vivo* reveals the importance of nucleosome occupancy in target site selection. *Genome Res.*, **16**, 1517–1528.
5. Almer, A., Rudolph, H., Hinnen, A. and Horz, W. (1986) Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements. *Embo J.*, **5**, 2689–2696.
6. Sekinger, E.A., Moqtaderi, Z. and Struhl, K. (2005) Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell*, **18**, 735–748.

7. Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M. and Iyer, V.R. (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.*, **6**, e65.
8. Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., Leproust, E.M., Hughes, T.R., Lieb, J.D., Widom, J. *et al.* (2008) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.
9. Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P. and Widom, J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
10. Yu, L. and Morse, R.H. (1999) Chromatin opening and transactivator potentiation by RAP1 in *Saccharomyces cerevisiae*. *Mol. Cell Biol.*, **19**, 5279–5288.
11. Whitehouse, I., Rando, O.J., Delrow, J. and Tsukiyama, T. (2007) Chromatin remodelling at promoters suppresses antisense transcription. *Nature*, **450**, 1031–1035.
12. Morse, R.H. (2003) Getting into chromatin: how do transcription factors get past the histones? *Biochem. Cell Biol.*, **81**, 101–112.
13. Hartley, P.D. and Madhani, H.D. (2009) Mechanisms that specify promoter nucleosome location and identity. *Cell*, **137**, 445–458.
14. Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
15. Tsai, H.K., Chou, M.Y., Shih, C.H., Huang, G.T., Chang, T.H. and Li, W.H. (2007) MYBS: a comprehensive web server for mining transcription factor binding sites in yeast. *Nucleic Acids Res.*, **35**, W221–W226.
16. Eriksson, P.R., Mendiratta, G., McLaughlin, N.B., Wolfsberg, T.G., Marino-Ramirez, L., Pompa, T.A., Jainerin, M., Landsman, D., Shen, C.H. and Clark, D.J. (2005) Global regulation by the yeast Spt10 protein is mediated through chromatin structure and the histone upstream activating sequence elements. *Mol. Cell Biol.*, **25**, 9127–9137.
17. Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
18. Dennis, G. Jr, Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.
19. Eden, E., Lipson, D., Yogeve, S. and Yakhini, Z. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.*, **3**, e39.
20. Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I. and Pugh, B.F. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, **18**, 1073–1083.
21. Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J. and Rando, O.J. (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, **309**, 626–630.
22. Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
23. Triolo, T. and Sternglanz, R. (1996) Role of interactions between the origin recognition complex and SIR1 in transcriptional silencing. *Nature*, **381**, 251–253.
24. Best, D.J. and Roberts, D.E. (1975) Algorithm AS 89: the upper tail probabilities of Spearman's Rho. *J. Royal Stat. Soc. C*, **24**, 377–379.
25. Westholm, J.O., Xu, F., Ronne, H. and Komorowski, J. (2008) Genome-scale study of the importance of binding site context for transcription factor binding and gene regulation. *BMC Bioinformatics*, **9**, 484.
26. Tharakaraman, K., Bodenreider, O., Landsman, D., Spouge, J.L. and Marino-Ramirez, L. (2008) The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. *Nucleic Acids Res.*, **36**, 2777–2786.
27. Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R. and Nislow, C. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.
28. Venters, B.J. and Pugh, B.F. (2009) A canonical promoter organization of the transcription machinery and its regulators in the *Saccharomyces* genome. *Genome Res.*, **19**, 360–371.
29. Lascaris, R.F., Groot, E., Hoen, P.B., Mager, W.H. and Planta, R.J. (2000) Different roles for abf1p and a T-rich promoter element in nucleosome organization of the yeast RPS28A gene. *Nucleic Acids Res.*, **28**, 1390–1396.
30. Chasman, D.I., Lue, N.F., Buchman, A.R., LaPointe, J.W., Lorch, Y. and Kornberg, R.D. (1990) A yeast protein that influences the chromatin structure of UASG and functions as a powerful auxiliary gene activator. *Genes Dev.*, **4**, 503–514.
31. Yarragudi, A., Miyake, T., Li, R. and Morse, R.H. (2004) Comparison of ABF1 and RAP1 in chromatin opening and transactivator potentiation in the budding yeast *Saccharomyces cerevisiae*. *Mol. Cell Biol.*, **24**, 9152–9164.
32. Badis, G., Chan, E.T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C.D., Gossett, A.J., Hasinoff, M.J., Warren, C.L. *et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell*, **32**, 878–887.
33. Angermayr, M., Oechsner, U. and Bandlow, W. (2003) Reb1p-dependent DNA bending effects nucleosome positioning and constitutive transcription at the yeast profilin promoter. *J. Biol. Chem.*, **278**, 17918–17926.
34. Devlin, C., Tice-Baldwin, K., Shore, D. and Arndt, K.T. (1991) RAP1 is required for BAS1/BAS2- and GCN4-dependent transcription of the yeast HIS4 gene. *Mol. Cell Biol.*, **11**, 3642–3651.
35. Goncalves, P.M., Griffioen, G., Minnee, R., Bosma, M., Kraakman, L.S., Mager, W.H. and Planta, R.J. (1995) Transcription activation of yeast ribosomal protein genes requires additional elements apart from binding sites for Abf1p or Rap1p. *Nucleic Acids Res.*, **23**, 1475–1480.
36. Shore, D. (1994) RAP1: a protean regulator in yeast. *Trends Genet.*, **10**, 408–412.
37. Moretti, P. and Shore, D. (2001) Multiple interactions in Sir protein recruitment by Rap1p at silencers and telomeres in yeast. *Mol. Cell Biol.*, **21**, 8082–8094.
38. Loo, S., Laurensen, P., Foss, M., Dillin, A. and Rine, J. (1995) Roles of ABF1, NPL3, and YCL54 in silencing in *Saccharomyces cerevisiae*. *Genetics*, **141**, 889–902.
39. Lieb, J.D., Liu, X., Botstein, D. and Brown, P.O. (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.*, **28**, 327–334.

A c-Myc regulatory subnetwork from human transposable element sequences†‡

Jianrong Wang,^a Nathan J. Bowen,^b Leonardo Mariño-Ramírez^{cd} and I. King Jordan^{*a}

Received 28th April 2009, Accepted 23rd June 2009

First published as an Advance Article on the web 21st July 2009

DOI: 10.1039/b908494k

Transposable elements (TEs) can donate regulatory sequences that help to control the expression of human genes. The oncogene c-Myc is a promiscuous transcription factor that is thought to regulate the expression of hundreds of genes. We evaluated the contribution of TEs to the c-Myc regulatory network by searching for c-Myc binding sites derived from TEs and by analyzing the expression and function of target genes with nearby TE-derived c-Myc binding sites. There are thousands of TE sequences in the human genome that are bound by c-Myc. A conservative analysis indicated that 816–4564 of these TEs contain canonical c-Myc binding site motifs. c-Myc binding sites are over-represented among sequences derived from the ancient TE families L2 and MIR, consistent with their preservation by purifying selection. Genes associated with TE-derived c-Myc binding sites are co-expressed with each other and with c-Myc. A number of these putative TE-derived c-Myc target genes are differentially expressed between Burkitt's lymphoma cells *versus* normal B cells and encode proteins with cancer-related functions. Despite several lines of evidence pointing to their regulation by c-Myc and relevance to cancer, the set of genes identified as TE-derived c-Myc targets does not significantly overlap with two previously characterized c-Myc target gene sets. These data point to a substantial contribution of TEs to the regulation of human genes by c-Myc. Genes that are regulated by TE-derived c-Myc binding sites appear to form a distinct c-Myc regulatory subnetwork.

Introduction

Almost half of the human genome sequence is made up of interspersed repeat sequences, which are remnants of formerly mobile transposable elements (TEs).^{1,2} These TE sequences have shaped the structure, function and evolution of their host genomes in a number of ways.^{3,4} For example, TEs are the source of a variety of regulatory sequences, including transcription factor binding sites (TFBS), alternative transcription start sites and small RNAs, that help to control the expression of host genes.⁵ The gene regulatory properties of TEs have received a great deal of attention in recent years, particularly since eukaryotic genome sequences and functional genomic datasets began accumulating over the last decade.

The ability of TEs to donate sequences that regulate nearby genes was first noticed in individual molecular genetic studies where regulatory elements were found to be located inside of repetitive sequence elements. In perhaps the first example of this kind of study, the sex-limited protein (Slp) encoding gene in mouse was shown to be regulated by androgen response elements located in the long terminal repeat sequence of an upstream endogenous retrovirus.⁶ An accumulation of such anecdotal cases was taken to support the possibility that TEs may have broad genome-scale effects on gene regulation.^{7,8} In the genomics era, three distinct classes of approaches have been taken to elucidate the regulatory contributions of TEs on the genome scale: (i) computational prediction of TE-derived regulatory sequences, (ii) identification of highly conserved TE sequences with comparative genomics and (iii) co-location of experimentally characterized regulatory sequences and TEs.

Computational analyses of TE sequences using position weight matrices that represent *cis*-regulatory sequence motifs have shown that TEs harbour numerous putative TFBS.^{9,10} These data, taken together with the genomic abundance of TEs, underscore their potential ability to regulate the expression of numerous host genes. A problem with this approach is that the *ab initio* prediction of *cis*-regulatory sequence motifs is prone to numerous false positives. To overcome this limitation, authors have used sequence shuffling, or simulation, to build null background sequence sets and then find TFBS that are over-represented among TE sequences relative to the background sets.¹⁰ Even with such a control for sequence composition in

^a School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA

E-mail: king.jordan@biology.gatech.edu, jwang64@gatech.edu

^b School of Biology and Ovarian Cancer Institute, Georgia Institute of Technology, Atlanta, GA 30332, USA. E-mail: bowen@gatech.edu

^c National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. E-mail: lmarino@ncbi.nlm.nih.gov

^d Computational Biology and Bioinformatics Unit, Bioindustry and Biotechnology Center, Corporación Colombiana de Investigación Agropecuaria – CORPOICA, Bogotá, Colombia

† This article is part of a *Molecular BioSystems* themed issue on Computational and Systems Biology.

‡ Electronic supplementary information (ESI) available: Supplementary figures, tables and human genome coordinates for TE-derived binding sites and co-located genes. See DOI: 10.1039/b908494k

place, it is still difficult to know which of these TE-derived TFBS may actually be functionally relevant in terms of regulating the expression of host genes. It can also be difficult to distinguish between sequences that regulate the expression of the element itself *versus* those that regulate nearby host genes.

Comparative genomics studies have been used to help identify TE sequences that are likely to encode functions for their host genomes. The rationale behind this approach is that conserved TE sequences have been preserved by purifying selection because of their functional, presumably regulatory, importance to the host organism.¹¹ The comparative genomics approach to the identification of TE-derived regulatory sequences was pioneered by Silva *et al.* who identified numerous ancient L2 and MIR intergenic TE sequences that were highly conserved among mammals and therefore likely to be functionally important.¹² Since that time, a number of studies have turned up thousands of conserved non-coding sequences that are derived from TEs.^{13–17} These findings indicate that a substantial fraction of TE sequences in mammalian genomes have been conserved by virtue of their functional (regulatory) relevance.¹⁸ However, this evolutionary approach to the identification of TE-derived regulatory sequences is overly conservative in some cases because it will not detect regulatory sequences that are derived from relatively recently inserted, or lineage-specific, TEs. Indeed, TEs are the most dynamic and rapidly evolving sequences in eukaryotic genomes, and most TE insertions are not shared between evolutionary lineages.¹⁹ Accordingly, it has been shown that numerous experimentally characterized TE-derived regulatory sequences are not conserved between species.^{19–21}

One of the most promising genome-scale approaches for the characterization of TE-derived regulatory sequences involves co-locating experimentally characterized regulatory elements and TE annotations on genomic sequences. This approach was first used on a relatively small scale by mapping the locations of hundreds of TFBS characterized in individual experiments to human TEs and then extrapolating to the entire human genome.²² This study suggested that thousands of human genes may be regulated by TE-derived regulatory sequences, but it was not possible to know whether this was actually the case. In order for the co-localization approach to really work on the genome-scale, high-throughput experimental data on the locations of regulatory sequences are needed. These data have become widely available in the last few years thanks to the invention of techniques like chromatin immunoprecipitation followed by microarray, ChIP-chip, or high-throughput sequencing, ChIP-Seq, analysis.²³ There are now hundreds-of-thousands of experimentally characterized TFBS that have been mapped to the human genome using these techniques, and recent studies have shown that many of these sites are derived from TEs.^{24,25} Many of these TE-derived TFBS are lineage-specific and may define recently evolved regulatory subnetworks that elaborate on previously existing networks as is the case for p53 binding sites derived from human endogenous retroviruses.²⁵

One particularly interesting transcription factor for which there is a human genome-wide map of binding sites is c-Myc.²⁶ c-Myc has been reported to regulate a large set of genes,^{27–29}

and it is considered an oncogene by virtue of its deregulation in a variety of cancers. For instance, c-Myc is markedly deregulated in lymphomas where it is over-expressed relative to normal B cells. A recent report evaluated the contribution of TEs to c-Myc binding sites on the human genome.²⁴ These authors found that c-Myc bound regions were not statistically enriched for co-localization with any particular TE family, and based on this observation concluded that c-Myc TFBS do not reside on repeats. However, our own preliminary data revealed that numerous c-Myc bound regions were in fact derived from human TE sequences, and we wanted to further explore the relationship between c-Myc binding and TEs to address this discrepancy.

Despite the lack of enrichment for c-Myc binding sites in a particular TE class or family observed previously, we found thousands of TE-derived c-Myc binding sites in the human genome using a conservative approach that integrated data from the experimental characterization of c-Myc bound regions with c-Myc binding site motif prediction. Gene expression and gene set enrichment analyses indicate that many of these TE-derived c-Myc binding sites are likely to be functionally relevant with respect to the regulation of human gene expression. However, most genes associated with TE-derived c-Myc binding sites do not correspond to genes previously characterized as targets of c-Myc regulation. This raises the possibility that TE-derived c-Myc targets define a distinct c-Myc regulatory subnetwork.

Results and discussion

TE-derived c-Myc binding sites

We integrated experimental data on c-Myc bound genomic sequences, probabilistic transcription factor binding site (TFBS) analysis and TE genome-annotations to identify TE-derived c-Myc binding sites in the human genome. The locations of c-Myc bound human genome sequences were previously determined using genome-wide chromatin immunoprecipitation (ChIP) and paired-end-ditag (PET) sequencing on P493 B cells.²⁶ We co-located these c-Myc bound human genome sequences with TE sequences annotated by RepeatMasker (<http://www.repeatmasker.org>). This analysis resulted in a set of 259 294 TE sequences co-located with c-Myc bound regions. The precise locations of TE-derived c-Myc binding sites were then determined by running the program Clover³⁰ on the c-Myc bound TE sequences. Clover was run using two c-Myc position–frequency matrices (Fig. S1, ESI†) with *P*-value thresholds of 0.01 and 0.001. This analysis resulted in a total of 4564 TE-derived c-Myc binding sites for $P \leq 0.01$ and 816 TE-derived c-Myc sites for $P \leq 0.001$. Thus, there is a substantial potential for human TE sequences to contribute to c-Myc gene regulatory networks. Here it should be noted that the use of Clover for the identification of specific c-Myc binding sites represents a conservative approach that eliminates many c-Myc bound human TE sequences that do not contain canonical c-Myc binding site sequence motifs. In fact, running Clover resulted in a two orders-of-magnitude reduction in the number of TE-derived c-Myc bound regions identified in the human genome.

Table 1 Number of TEs that contain c-Myc binding sites for each TE class/family

TE class/family ^a	Observed number ^b	Observed percent ^c (%)	Expected percent ^d (%)
L1	940	20.60	21.9
L2	546	11.96	9.7
LINE other	47	1.03	1.6
Alu	994	21.78	28.1
MIR	733	16.06	13.9
SINE other	27	0.59	0.1
DNA	411	9.01	9.3
LTR	866	18.97	15.5
Total	4564	100.00	100.0

^a Name of TE classes or families. ^b Observed number of TEs in each class/family. ^c Observed percent: the observed number of TEs in each class/family divided by the total observed number (4564) of TEs containing c-Myc binding sites. ^d Expected percent: the total number of TEs in each class/family in human genome divided by the total number of all TEs in human genome.

While this approach may result in the loss of some *bona fide* TE-derived c-Myc binding sites, it also yields increased confidence in the functional relevance of the smaller set of TE-derived sites we identified.

In order to evaluate the contribution of distinct TEs to c-Myc binding sites, we divided human TEs into 8 classes/families based on the Repbase database³¹ classification system: L1, L2, LINE other, Alu, MIR, SINE other, DNA and LTR. The observed numbers of individual TE insertions with c-Myc binding sites for each class/family are shown in Table 1 ($P \leq 0.01$) and Table S1, ESI† ($P \leq 0.001$), and a comparison of the observed *versus* expected percentages for each TE class/family are shown in Fig. 1A ($P \leq 0.01$) and Fig. S2, ESI† ($P \leq 0.001$). Members of the abundant L1 and Alu element families have lower observed than expected percentages, while L2 and MIR elements have higher than expected percentages. The relative ages of these families can be estimated by calculating the sequence divergence between individual elements and subfamily consensus sequences; younger elements have lower divergence since they inserted in the genome more recently. L1s and Alus are younger element families, many of which are primates-specific, whereas L2 and MIR are more ancient families that radiated early in mammalian evolution (Fig. 1B). In other words, relatively older TE families contribute

more c-Myc binding sites than expected based on their percentage in the genome, whereas younger families contribute fewer c-Myc binding sites than expected. A similar pattern was found in a recent study that analyzed experimentally characterized human TE-derived binding sites from numerous distinct transcription factors.²¹ The enrichment of c-Myc TFBS in more ancient TEs is consistent with the notion that these sequences have been conserved in the genome by purifying selection based on their functional relevance.¹² Nevertheless, Alu elements show the highest number of c-Myc binding sites, since they are the most numerous elements in the genome. TFBS derived from relatively young, even polymorphic in some cases, elements like Alu are of interest since they may impart lineage- or condition-specific regulatory properties on nearby genes.^{18–21,25} We explore this possibility later in the manuscript.

Regulatory effects of TE-derived c-Myc binding sites

In order to evaluate the potential regulatory effects of TE-derived c-Myc binding sites, we mapped the TE-derived sites to the vicinity of human genes and analyzed these genes' tissue-specific expression patterns. Human genes with TE-derived c-Myc binding sites within ± 10 kb were considered as potential c-Myc regulated target genes. This resulted in a

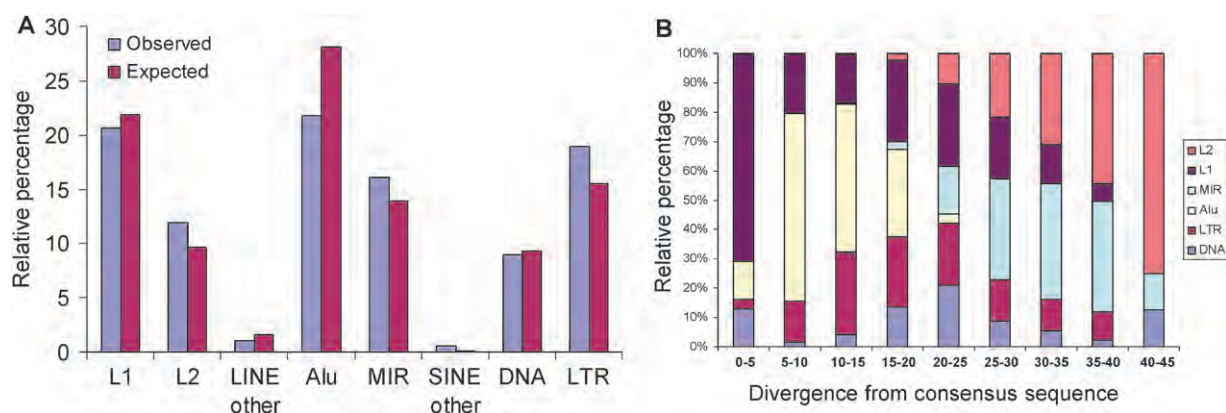


Fig. 1 Family origins and relative ages for human TEs bound by c-Myc. Observed *versus* expected percentages of c-Myc binding sites derived from different TE classes/families. (A) The observed percentages (blue) of TEs containing c-Myc binding sites in each TE class/family are plotted along with the expected percentages (maroon) of TEs in each class/family based on their background percentages in the human genome. (B) Percent divergence from subfamily consensus sequences for human TEs that are bound by c-Myc. The relative percentages of each of the six TE families are shown for each percent divergence bin. Younger elements have lower divergence from their consensus sequences, and older elements have higher divergence.

Table 2 Pearson correlation coefficients (PCC) of gene expression within each target gene class

Target gene class ^a	Number of probes ^b	Average PCC ^c	Z score ^d	P-value ^e
L1	357	0.027	31.17	3.85×10^{-213}
L2	297	0.031	29.63	7.96×10^{-193}
LINE other	20	0.033	2.29	0.022
Alu	550	0.044	73.06	0
MIR	390	0.021	26.10	4.64×10^{-150}
SINE other	17	0.055	2.32	0.021
DNA	205	0.028	17.57	4.04×10^{-69}
LTR	257	0.021	16.63	4.13×10^{-62}

^a Name of TE classes or families. ^b Number of Affymetrix probes corresponding to genes with c-Myc binding sites derived from TE of specific classes/families. ^c Average of Pearson correlation coefficients (PCC) of each pair of probes within specific TE classes/families. ^d Z-transformation of PCC values. ^e P-values indicate the significance levels of Z scores.

total set of 1550 human genes with proximal TE-derived c-Myc binding sites. The expression patterns of these putative target genes over 79 human tissues and cell lines were compared to each other, and to the expression patterns of c-Myc, using the Novartis human gene expression atlas of Affymetrix microarray data.³²

For each class/family of TEs, the expression patterns of all putative c-Myc target genes were compared using the Pearson correlation coefficient (PCC). Table 2 shows the number of target gene Affymetrix probes for each TE class/family along with the average PCCs, Z scores and P-values. All 8 TE classes/families have sets of putative c-Myc target genes that are positively and significantly co-expressed, on average, across human tissues. Target genes with Alu-derived c-Myc binding sites, the most numerous class, show the highest levels of average co-expression and the greatest statistical significance. It should be noted that while the average co-expression levels for the distinct TE class/family target gene sets are all positive and, for the most part, highly statistically significant, the average PCC values are still quite low (*i.e.* close to 0). This suggests that while there is certainly an enrichment for co-expressed gene pairs among the TE-derived c-Myc target genes, the total set of target genes for each class has a broad range of tissue-specific expression patterns. This is consistent with the fact that genes with proximal TE-derived c-Myc binding sites are also likely to be regulated by additional transcription factors as well as different classes of regulators such as epigenetic modifications and/or small RNAs.

In order to further explore the relationship between human gene expression and the presence of TE-derived c-Myc binding sites, tissue-specific expression levels of putative target genes were compared to the expression of the regulator c-Myc. This allowed us to more directly investigate whether those target genes are actually regulated by c-Myc. To do this, we calculated the target genes' average expression levels in each tissue and compared them with the c-Myc expression data by calculating pairwise PCCs across tissues between the TE classes/families and c-Myc. The results of the PCC analysis are shown in Table 3, and the average expression levels for TE classes/families and c-Myc across 79 tissues are shown in Fig. 2. 7 out of 8 TE class/family target gene sets show statistically significant co-expression with c-Myc. Furthermore, for these 7 TE classes/families, the average PCC values between the putative target genes with TE-derived binding sites and c-Myc are an order of magnitude greater (Table 3)

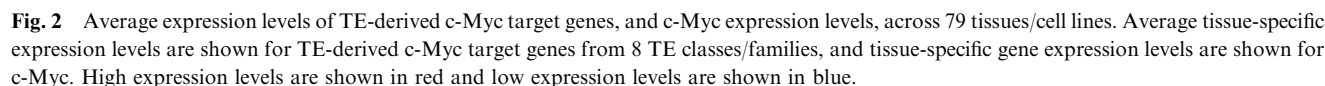
Table 3 Pearson correlation coefficients (PCC) between expression levels of TE-derived target genes and c-Myc

Target gene class ^a	PCC ^b	<i>t</i> ^c	P-value ^d
L1	0.37	3.45	9.18×10^{-04}
L2	0.35	3.28	1.57×10^{-03}
LINE other	-0.10	-0.87	0.39
Alu	0.48	4.79	7.95×10^{-06}
MIR	0.41	3.93	1.86×10^{-04}
SINE other	0.62	7.00	8.17×10^{-10}
DNA	0.34	3.14	2.41×10^{-03}
LTR	0.32	2.94	4.36×10^{-03}

^a Name of TE classes or families. ^b Pearson correlation coefficients (PCC) between the average tissue-specific expression levels of all target genes with a TE class/family and c-Myc. ^c PCC transformed into *t*-values by $t = \text{PCC} \times \sqrt{\text{df}/(1-\text{PCC}^2)}$ where $\text{df} = 77$. ^d P-values indicate the significance levels of *t* scores (following Student's *t* distribution).

than the average PCC values among all pairs of target genes (Table 2). This indicates that the target genes' tissue-specific expression patterns are distributed around the expression pattern of c-Myc in such a way as to be more similar to c-Myc, on average, than they are to each other. This can be visually appreciated by comparing the average tissue-specific expression levels of the TE class/family target genes to the expression pattern of c-Myc (Fig. 2). Target genes with TE-derived c-Myc binding sites are clearly more highly expressed, on average, in the same tissues where c-Myc is also highly expressed. The most striking cases of c-Myc-to-target gene co-expression can be seen for both normal and cancerous T cells and B cells, including CD4⁺ and CD8⁺ T cells, CD19⁺ B cells and several lymphoma and leukemia cell lines (Fig. 2).

We performed a permutation test to more precisely identify the specific tissues where both c-Myc and the target genes with TE-derived c-Myc binding sites are over-expressed. To do this, the average tissue-specific expression levels of all target genes were computed and compared to 1000 randomly permuted (over the same gene set) tissue-specific average expression level vectors. The same analysis was done using c-Myc tissue-specific expression levels as the test set. For each tissue, the observed test set average, or c-Myc, expression level was then compared to the distribution of permuted values. There are 22 significantly ($P < 0.05$) over-expressed tissues among the TE c-Myc binding site target genes including the aforementioned normal and cancerous T and B cells as well as several brain



A TE-specific c-Myc regulatory network

c-Myc target gene sets

- A** - Basso *et. al* $n=2,063$
- B** - Zeller *et. al* $n=1,617$
- C** - TE-derived targets $n=1,550$

Venn Diagram Data:

Region	Count
A only	1629
B only	1263
C only	1401
A ∩ B	412
A ∩ C	67
B ∩ C	60
A ∩ B ∩ C	22

Contingency Table:

	A	B	C	A ∪ B	A ∩ B	A ∩ C	B ∩ C
A		434	89				
B	1.5×10^{-10}		82				
C	0.99998	0.99257		149	22	67	60
A ∪ B			0.99995				
A ∩ B			0.91044				
A ∩ C			0.99991				
B ∩ C			0.98592				

network. This interpretation is consistent with previously published results indicating that TEs can provide lineage-specific regulatory sequences.^{18–21,25}

In order to try and discriminate between these two possible scenarios, (i) functional irrelevance of the TE-derived c-Myc binding sites *versus* (ii) a TE-specific c-Myc regulatory network, we evaluated the overlap of the TE-derived c-Myc target gene set with a series of gene set collections from the molecular signatures database (MSigDB) (<http://www.broad.mit.edu/gsea/msigdb/index.jsp>). The MSigDB gene sets represent groups of genes with similar features or properties such as co-regulated genes, genes with similar *cis*-regulatory motifs and genes with similar gene ontology (GO) functional annotations.³³ Thus, gene set enrichment analysis with the MSigDB can be used to evaluate whether the TE-derived c-Myc target genes have similar biological functions or regulation. The TE-derived c-Myc target genes were broken down into class/family-specific sets and run against MSigDB.

This analysis resulted in numerous statistically significant gene set enrichments (Table S3, ESI†), the most relevant of which include a number of cancer related gene modules. These data suggest that many of the TE-derived c-Myc target genes are functionally related and associated with cancer. For instance, c-Myc target genes with L2-derived binding sites are enriched for a cluster of genes with expression patterns indicative of lymphoma and immune response, based on their tissue-specific expression levels. Both MIR and LTR elements donate c-Myc binding sites to genes classified as being involved in B cell lymphoma *via* so-called clinical annotations, which associate microarrays with known clinical attributes. In other words, TE-derived c-Myc target genes that are from different families, and are identified with different methodologies, converge on genes that function in B cells and in cancer. In addition, DNA element-derived c-Myc target genes are enriched for genes involved in the MAPK signalling pathway, which regulates cellular response to growth factors and mediates the action of many oncogenes.

Differential expression in Burkitt's lymphoma *versus* normal B cell

c-Myc is a well known oncogene that is over-expressed in a number of different cancers, particularly lymphomas.²⁸ In light of its role in cancer, we asked whether TE-derived c-Myc target genes showed differential expression between cancer and normal cells. To do this, we used a microarray gene expression dataset, from the Oncomine database, comparing Burkitt's lymphoma ($n = 31$) *versus* normal B cell

($n = 25$).²⁷ We identified 53 TE-derived c-Myc target genes that show statistically significant ($P < 0.05$) differential expression between normal and cancer (Fig. 4); 16 of the c-Myc binding sites that map to these genes are derived from Alu elements. c-Myc is also known to be over-expressed in Burkitt's lymphoma cells, and we calculated the PCC across the 56 cancer and normal cell lines for these 53 genes' expression data with c-Myc's to see if the differentially expressed target genes are co-regulated with c-Myc (Table S4, ESI†). There are 32 TE-derived c-Myc target genes that show positive correlations ($0.23 \leq \text{PCC} \leq 0.80$) with c-Myc and 21 target genes with negative correlations ($-0.66 \leq \text{PCC} \leq -0.38$); all PCC are statistically significant ($P < 0.05$). These data indicate that TE-derived c-Myc binding sites contribute to the cancer-related expression of c-Myc regulated genes. TE-derived c-Myc target genes are both up-regulated and down-regulated in cancer, while c-Myc is over-expressed in lymphoma relative to normal B cells. This finding may be attributed to the fact that c-Myc can both positively and negatively regulate the expression of its target genes.²⁸ The fact that the majority of correlations are positive is consistent with our results showing the overall average positive correlation between TE-derived c-Myc target genes and c-Myc (Table 3 and Fig. 2).

In order to further evaluate the function of these differentially expressed genes, gene set enrichment analysis was performed on the set of 53 TE-derived c-Myc target genes that are deregulated in Burkitt's lymphoma. To do this, the genes were sorted according to the TE class/family of their c-Myc binding sites and each set was evaluated against the MSigDB gene sets. A number of statistically significant enrichments for

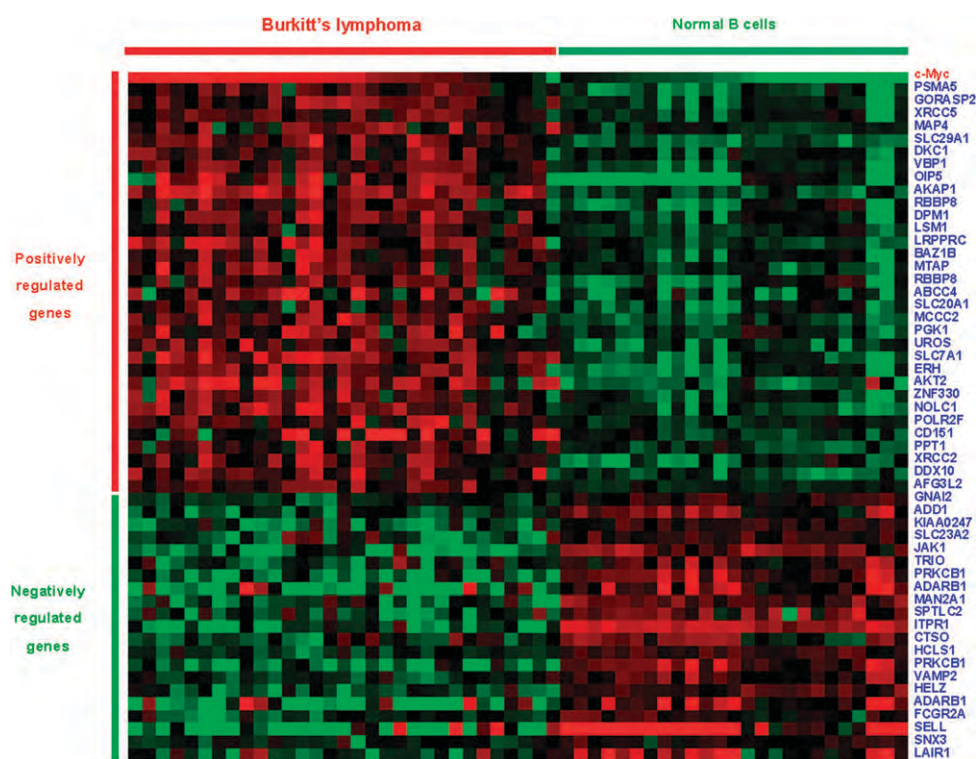


Fig. 4 Differential expression of TE-derived c-Myc target genes in Burkitt's lymphoma *versus* normal B cells. Each row shows the expression levels of a gene in Burkitt's lymphoma cells ($n = 31$ on the left) and normal cells ($n = 25$ on the right); over-expression is shown in red and under-expression in green.

cancer-related gene sets were detected, particularly for MIR and L1 elements (Table S5, ESI†). For instance, the genes encoding ITPR1 and AKT2 bear MIR-derived c-Myc binding sites and show up in several enriched gene sets including members of the B cell antigen receptor signalling pathway genes and the gene set related to the PIP3 signalling pathway in B lymphocytes. For instance, AKT genes encode serine–threonine protein kinases that promote cell proliferation by phosphorylating targets that lead to the activation of the anti-apoptotic transcription factor NF- κ B. ITPR1 encodes an intracellular channel that mediates release of calcium from the endoplasmic reticulum, which can also lead to cell proliferation *via* stimulation of the CALML6 protein upstream in the calcium signalling pathway. In addition, the genes LRPPC and PRKCB1 both have L1-derived c-Myc binding sites and are known to be deregulated in B cell lymphoma.

Alu elements are the single most abundant class/family of TEs that provide c-Myc binding sites to human genes, and Alu-derived c-Myc binding sites are also over-represented among the set of target genes differentially expressed between Burkitt's lymphoma and cancer. As alluded to previously, we were particularly interested in Alu elements since they have inserted relatively recently in the human genome, are potentially polymorphic, and have a known role in several cancers.³⁴ We investigated the Alu-derived c-Myc target genes shown to be differentially regulated between Burkitt's lymphoma *versus* normal B cells and found a small set of Alu c-Myc target genes that were tightly coherent with respect to several different characteristics (Fig. 5). These genes all have Alu-derived c-Myc binding sites that are located around the 5' transcription start site, three of which are located within the proximal ± 2 kb promoter region (Fig. 5A). All of these genes are up-regulated in Burkitt's lymphoma and positively correlated with c-Myc expression (Table 4 and Fig. 5B). The specific c-Myc binding

sites in these Alu sequences are all derived from one particular location in the element suggesting that the c-Myc TFBS evolved in an ancestral sequence and was distributed by transposition, as opposed to evolving *in situ* after the elements inserted (Fig. 5C and Fig. S4, ESI†). Two of the five genes have c-Myc binding sites derived from AluSg subfamily sequences and the other three have c-Myc binding sites derived from the AluSx subfamily. AluSg and AluSx are particularly young Alu subfamilies that are polymorphic (*i.e.* show insertion site differences) among human populations.³⁵ It is possible that polymorphic Alu elements change the regulatory network of c-Myc between individual humans and/or between cell types. Furthermore, if a gene is brought under the control of c-Myc by an Alu insertion it could lead to changes in expression of that gene associated with oncogenesis. These recently evolved Alu-derived c-Myc binding sites exemplify TE contributions to a specific c-Myc subnetwork, consistent with our characterization of numerous novel c-Myc target genes that are associated with TE-derived binding sites.

Materials and methods

Identification of TE-derived c-Myc binding sites

The locations of experimentally characterized c-Myc bound regions, characterized previously by genome-wide chromatin immunoprecipitation (ChIP) and paired-end-tag (PET) sequencing on P493 B cells,²⁶ were taken from the GIS ChIP-PET track in UCSC genome browser (<http://www.genome.ucsc.edu/>).³⁶ The positions of TEs were taken from the RepeatMasker track in UCSC genome browser. TE and c-Myc bound regions were co-localized using the UCSC table browser tool.³⁷ TE-derived c-Myc bound regions were analyzed with the program Clover,³⁰ using two c-Myc binding site motif

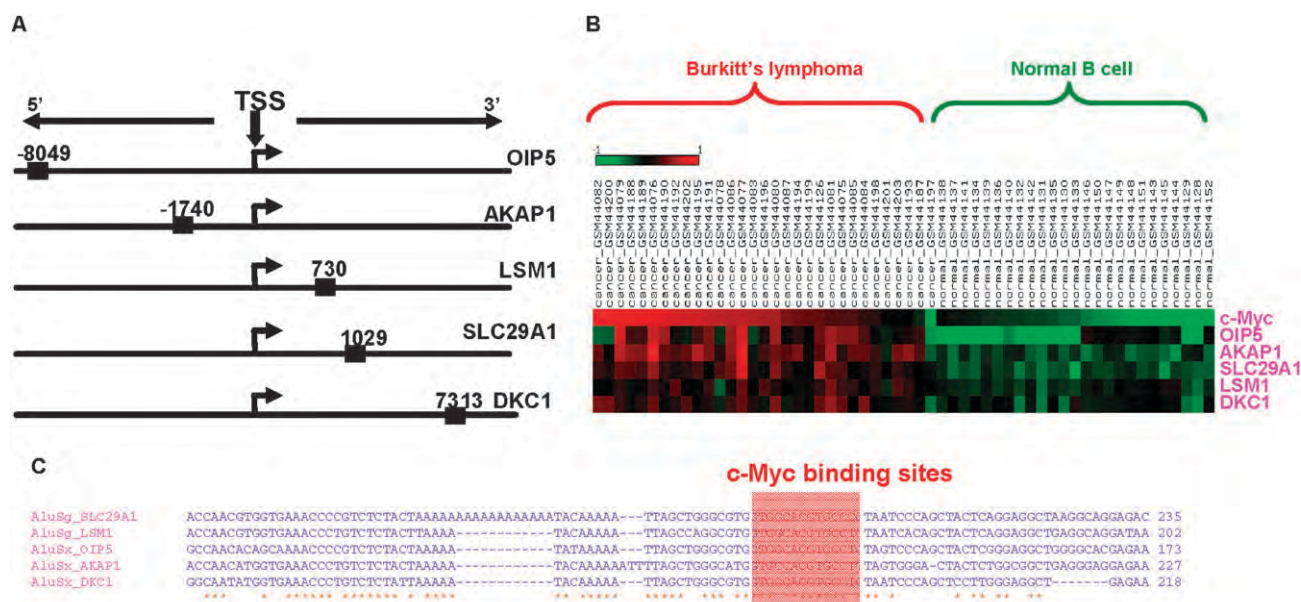


Fig. 5 Target genes with Alu-derived c-Myc binding sites. (A) Approximate illustration of relative positions of Alu-derived c-Myc binding sites compared with target genes' transcriptional start sites (TSS). (B) Differential expression of Alu-derived c-Myc target genes, and c-Myc, in Burkitt's lymphoma *versus* normal B cells. (C) Multiple sequence alignment of the Alu element insertions with c-Myc binding site locations indicated.

Table 4 Differential expression of Alu-derived c-Myc target genes

Gene symbol	Differential expression (<i>t</i> -value) ^a	Differential expression (<i>P</i> -value) ^b	Correlation with c-Myc ^c	<i>P</i> -value of correlation ^d
SLC29A1	11.98	1.4×10^{-16}	0.77	1.92×10^{-12}
LSM1	6.54	2.3×10^{-8}	0.50	4.70×10^{-5}
OIP5	5.78	1.9×10^{-6}	0.48	9.39×10^{-5}
AKAP1	11.21	1×10^{-14}	0.79	1.42×10^{-13}
DKC1	5.45	1.3×10^{-6}	0.64	5.27×10^{-8}

^a Gene's differential expression in Burkitt's lymphoma cells *versus* normal B cells. *T*-values computed by the Student's *t*-test. ^b Significance levels (*P*-values) of the differential expression. ^c Pearson correlation coefficients between cancer *versus* normal expression of Alu-derived c-Myc target genes and c-Myc. ^d Significance levels (*P*-values) of the correlation.

position–frequency matrices from the TRANSFAC database³⁸ (V\$MYC_01 and V\$MYC_02 see Fig. S1, ESI†) to precisely locate c-Myc binding sites. Clover uses non-parametric approach with 1000 randomizations of the search sequence to generate a score and associated *P*-value. Clover was run using a conservative score threshold of 6 with two *P*-value thresholds $P < 0.01$ and $P < 0.001$.

Human TE sequences were divided into 8 classes/families using the Repeat classification system³¹ implemented with RepeatMasker: L1, L2, LINE-other (LINE elements excluding L1 and L2), Alu, MIR, SINE-other (SINE elements excluding Alu and MIR), DNA and LTR. Alu elements were further divided into subfamilies and members of individual subfamilies bound by c-Myc were aligned using ClustalW³⁹ to identify the relative locations of c-Myc binding sites.

Analysis of TE-derived c-Myc target genes

Human Refseq⁴⁰ genes were identified as putative TE-derived c-Myc regulatory targets if they had TE-derived c-Myc binding sites within 10 kb of the gene boundaries. Microarray gene expression data were taken from the Novartis mammalian gene expression atlas version 2 (GNF2),³² and Affymetrix probes from GNF2 were mapped to TE-derived c-Myc target genes using the UCSC genome browser annotations. Co-expression among TE-derived c-Myc target genes, and between target genes and c-Myc, was evaluated by calculating Pearson correlation coefficients (PCC) between pairs of genes across 79 different tissues or cell lines. Statistical significance levels (*P*-values) of PCC values, and averages, were computed using the *Z* transformation. A permutation test was used to identify sets of tissues that are over-expressed for c-Myc and among all TE-derived c-Myc target genes. To do this, tissue-specific gene expression vectors were randomly shuffled for each gene and average tissue-specific expression values were calculated for all randomly shuffled genes. 1000 sets of average tissue-specific expression values were used to compute null background expression level distributions for each tissue against which the observed values were compared. All *P*-values were corrected for multiple tests using the Benjamini–Hochberg false discovery rate.

Differential expression of target genes in cancer *versus* normal cells

TE-derived c-Myc target genes were mapped to the Burkitt's lymphoma and normal B cell microarray dataset compiled by Basso *et al.*²⁷ The Oncomine database⁴¹ was used to select genes from this dataset that were determined to be differentially

expressed between cancer (Burkitt's lymphoma $n = 31$) *versus* normal B cells ($n = 25$) using the Student's *t*-test. Co-expression values between these differentially expressed TE-derived c-Myc target genes and c-Myc, across the 56 cancer and normal B cell lines, were computed using the PCC as described previously.

Gene set enrichment and c-Myc target gene analyses

Sets of TE-derived c-Myc target genes for each TE class/family were searched against a series of gene set collections from the molecular signatures database (MSigDB)³³ to evaluate their shared functional and/or regulatory features. The extent and significance of the overlaps between the set of TE-derived c-Myc target genes identified here and two previously characterized c-Myc target gene sets were evaluated using the hypergeometric distribution:

$$P(X \geq k) = \sum_{i=k}^N \frac{\binom{n}{i} \binom{m}{N-i}}{\binom{n+m}{N}}$$

where k = number of overlapping target genes, N = number of TE-derived c-Myc target genes, n = number of previously characterized c-Myc target genes, and m = human genes not previously characterized as c-Myc targets.

Conclusions

Recently, Bourque *et al.* analyzed the ability of human TEs to provide transcription factor binding sites genome-wide.²⁴ They considered high-throughput binding site data for seven transcription factors, including c-Myc analyzed here, and concluded that five of these transcription factors bind to distinct families of human TEs. However, c-Myc was not one of the families identified in their study to bind to human TEs. This can be attributed to the enrichment criteria used to characterize transcription factors as binding human TEs. Specifically, they only considered transcription factors that bind to families of TEs with higher than expected frequency based on the abundance of the TE in the genome. This approach makes sense from a quantitative perspective, but it may be overly conservative if it misses *bona fide* functional transcription factor binding sites derived from TEs. We found that hundreds-of-thousands of human TEs have experimental evidence of being bound by c-Myc. Furthermore, many of these TE sequences harbor canonical c-Myc binding site sequence motifs, suggesting that the binding of c-Myc to the

elements is not spurious. In addition, our own functional analysis of human genes with proximal TE-derived c-Myc binding sites suggests that many of these sites may indeed be functional with respect to mediating gene regulation by c-Myc. However, definitive proof of such function will have to await experimental characterization. Hopefully, the list of gene targets and TE-derived c-Myc binding sites uncovered by our analysis can be used to stimulate investigation of the regulatory properties of human TEs.

TE sequences in the human genome provide thousands of c-Myc binding sites, and genes that bear nearby TE-derived sites show evidence for regulation by c-Myc. TE-mediated regulation of human genes by c-Myc includes changes in expression that are characteristic of the difference between cancer *versus* normal B cells, and TE-derived target genes encode proteins with cancer-related functions. Nevertheless, the TE-derived c-Myc target genes identified in this study do not overlap, for the most part, with previously characterized c-Myc target genes. This suggests that expansion of TE sequences may provide a mechanism for the emergence of distinct lineage-specific regulatory subnetworks.²⁵

Acknowledgements

I. K. J. and J. W. were supported by the School of Biology, Georgia Institute of Technology. I. K. J. was supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). N. J. B. was supported by the Ovarian Cancer Institute Laboratory, School of Biology, Georgia Institute of Technology. This research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI. LMR was supported by Corporación Colombiana de Investigación Agropecuaria—CORPOICA.

References

- 1 E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford and J. Howland, *et al.*, *Nature*, 2001, **409**, 860–921.
- 2 J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson and J. R. Wortman, *et al.*, *Science*, 2001, **291**, 1304–1351.
- 3 C. Biemont and C. Vieira, *Nature*, 2006, **443**, 521–524.
- 4 H. H. Kazazian, Jr., *Science*, 2004, **303**, 1626–1632.
- 5 C. Feschotte, *Nat. Rev. Genet.*, 2008, **9**, 397–405.
- 6 J. B. Stavenhagen and D. M. Robins, *Cell*, 1988, **55**, 247–254.
- 7 R. J. Britten, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 9374–9377.
- 8 R. J. Britten, *Gene*, 1997, **205**, 177–182.
- 9 R. Shankar, D. Grover, S. K. Brahmachari and M. Mukerji, *BMC Evol. Biol.*, 2004, **4**, 37.
- 10 B. G. Thornburg, V. Gotea and W. Makalowski, *Gene*, 2006, **365**, 104–110.
- 11 D. J. Witherspoon, T. G. Doak, K. R. Williams, A. Seegmiller, J. Seger and G. Herrick, *Mol. Biol. Evol.*, 1997, **14**, 696–706.
- 12 J. C. Silva, S. A. Shabalina, D. G. Harris, J. L. Spouge and A. S. Kondrashov, *Genet. Res.*, 2003, **82**, 1–18.
- 13 G. Bejerano, C. B. Lowe, N. Ahituv, B. King, A. Siepel, S. R. Salama, E. M. Rubin, W. J. Kent and D. Haussler, *Nature*, 2006, **441**, 87–90.
- 14 M. Kamal, X. Xie and E. S. Lander, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 2740–2745.
- 15 H. Nishihara, A. F. Smit and N. Okada, *Genome Res.*, 2006, **16**, 864–874.
- 16 A. M. Santangelo, F. S. de Souza, L. F. Franchini, V. F. Bumashny, M. J. Low and M. Rubinstein, *PLoS Genet.*, 2007, **3**, 1813–1826.
- 17 X. Xie, M. Kamal and E. S. Lander, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 11659–11664.
- 18 T. S. Mikkelsen, M. J. Wakefield, B. Aken, C. T. Amemiya, J. L. Chang, S. Duke, M. Garber, A. J. Gentles, L. Goodstadt, A. Heger, J. Jurka, M. Kamal, E. Mauceli, S. M. Searle and T. Sharpe, *et al.*, *Nature*, 2007, **447**, 167–177.
- 19 L. Marino-Ramirez, K. C. Lewis, D. Landsman and I. K. Jordan, *Cytogenet. Genome Res.*, 2005, **110**, 333–341.
- 20 L. Marino-Ramirez and I. K. Jordan, *Biol. Direct*, 2006, **1**, 20.
- 21 N. Polavarapu, L. Marino-Ramirez, D. Landsman, J. F. McDonald and I. K. Jordan, *BMC Genomics*, 2008, **9**, 226.
- 22 I. K. Jordan, I. B. Rogozin, G. V. Glazko and E. V. Koonin, *Trends Genet.*, 2003, **19**, 68–72.
- 23 G. M. Euskirchen, J. S. Rozowsky, C. L. Wei, W. H. Lee, Z. D. Zhang, S. Hartman, O. Emanuelsson, V. Stolc, S. Weissman, M. B. Gerstein, Y. Ruan and M. Snyder, *Genome Res.*, 2007, **17**, 898–909.
- 24 G. Bourque, B. Leong, V. B. Vega, X. Chen, Y. L. Lee, K. G. Srinivasan, J. L. Chew, Y. Ruan, C. L. Wei, H. H. Ng and E. T. Liu, *Genome Res.*, 2008, **18**, 1752–1762.
- 25 T. Wang, J. Zeng, C. B. Lowe, R. G. Sellers, S. R. Salama, M. Yang, S. M. Burgess, R. K. Brachmann and D. Haussler, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 18613–18618.
- 26 K. I. Zeller, X. Zhao, C. W. Lee, K. P. Chiu, F. Yao, J. T. Yustein, H. S. Ooi, Y. L. Orlov, A. Shahab, H. C. Yong, Y. Fu, Z. Weng, V. A. Kuznetsov, W. K. Sung and Y. Ruan, *et al.*, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 17834–17839.
- 27 K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera and A. Califano, *Nat. Genet.*, 2005, **37**, 382–390.
- 28 S. Pelengaris, M. Khan and G. Evan, *Nat. Rev. Cancer*, 2002, **2**, 764–776.
- 29 K. I. Zeller, A. G. Jegga, B. J. Aronow, K. A. O'Donnell and C. V. Dang, *Genome Biology*, 2003, **4**, R69.
- 30 M. C. Frith, Y. Fu, L. Yu, J. F. Chen, U. Hansen and Z. Weng, *Nucleic Acids Res.*, 2004, **32**, 1372–1381.
- 31 O. Kohany, A. J. Gentles, L. Hankus and J. Jurka, *BMC Bioinformatics*, 2006, **7**, 474.
- 32 A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker and J. B. Hogenesch, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 6062–6067.
- 33 A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15545–15550.
- 34 P. L. Deininger and M. A. Batzer, *Mol. Genet. Metab.*, 1999, **67**, 183–193.
- 35 J. Wang, L. Song, M. K. Gonder, S. Azrak, D. A. Ray, M. A. Batzer, S. A. Tishkoff and P. Liang, *Gene*, 2006, **365**, 11–20.
- 36 W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler and D. Haussler, *Genome Res.*, 2002, **12**, 996–1006.
- 37 D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler and W. J. Kent, *Nucleic Acids Res.*, 2004, **32**, D493–D496.
- 38 V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel and A. E. Kel, *et al.*, *Nucleic Acids Res.*, 2006, **34**, D108–D110.
- 39 J. D. Thompson, D. G. Higgins and T. J. Gibson, *Nucleic Acids Res.*, 1994, **22**, 4673–4680.
- 40 D. R. Maglott, K. S. Katz, H. Sicotte and K. D. Pruitt, *Nucleic Acids Res.*, 2000, **28**, 126–128.
- 41 D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey and A. M. Chinnaiyan, *Neoplasia*, 2004, **6**, 1–6.



Repetitive DNA elements, nucleosome binding and human gene expression

Ahsan Huda^a, Leonardo Mariño-Ramírez^{b,c}, David Landsman^b, I. King Jordan^{a,*}

^a School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30332, USA

^b National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

^c Computational Biology and Bioinformatics Unit, Biotechnology and Bioindustry Center, Corporacion Colombiana de Investigacion Agropecuaria - CORPOICA, Km. 14 Via a Mosquera, Bogota, Colombia

ARTICLE INFO

Article history:

Received 22 January 2009

Accepted 23 January 2009

Available online 5 February 2009

Received by M. Batzer

Keywords:

Transposable elements

Nucleosome binding

Epigenetics

Simple sequence repeats

Gene regulation

Promoter architecture

Human genome

ABSTRACT

We evaluated the epigenetic contributions of repetitive DNA elements to human gene regulation. Human proximal promoter sequences show distinct distributions of transposable elements (TEs) and simple sequence repeats (SSRs). TEs are enriched distal from transcriptional start sites (TSSs) and their frequency decreases closer to TSSs, being largely absent from the core promoter region. SSRs, on the other hand, are found at low frequency distal to the TSS and then increase in frequency starting ~150 bp upstream of the TSS. The peak of SSR density is centered around the –35 bp position where the basal transcriptional machinery assembles. These trends in repetitive sequence distribution are strongly correlated, positively for TEs and negatively for SSRs, with relative nucleosome binding affinities along the promoters. Nucleosomes bind with highest probability distal from the TSS and the nucleosome binding affinity steadily decreases reaching its nadir just upstream of the TSS at the same point where SSR frequency is at its highest. Promoters that are enriched for TEs are more highly and broadly expressed, on average, than promoters that are devoid of TEs. In addition, promoters that have similar repetitive DNA profiles regulate genes that have more similar expression patterns and encode proteins with more similar functions than promoters that differ with respect to their repetitive DNA. Furthermore, distinct repetitive DNA promoter profiles are correlated with tissue-specific patterns of expression. These observations indicate that repetitive DNA elements mediate chromatin accessibility in proximal promoter regions and the repeat content of promoters is relevant to both gene expression and function.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

The prevalence of repetitive DNA sequences in mammalian genomes has been appreciated since the classic re-association kinetic (COT-curve) experiments of the late nineteen-sixties (Britten and Kohne, 1968). The completion of the human genome projects at the turn of the millennium further underscored the extent to which the human genome sequence is made up of repetitive DNA elements (Lander et al., 2001; Venter et al., 2001). There are several distinct categories of repetitive sequence elements in the human genome. Interspersed repeat sequences, also known as transposable elements (TEs), make up at least 45% of the euchromatic genome sequence, and novel human TE families continue to be discovered and characterized (Wang et al., 2005; Nishihara et al., 2006). Simple sequence repeats (SSRs) consist of tandem repeats of exact or nearly exact units of length k (k -mers), with $k = 1$ –13 corresponding to microsatellites and $k = 1$ –500 for minisatellites. Analysis of the human genome sequence showed that ~3% of the euchromatic sequence was made up of SSRs, and both SSRs and TEs are thought to be

far more abundant in heterochromatin. Segmental duplications of 1–200 kb were initially shown to account for ~3% of the human genome sequence (Lander et al., 2001), and more recent results reveal that copy number variants populate the genome to an even greater extent (Kidd et al., 2008).

The evolutionary significance and the functional role that repetitive genomic elements, TEs in particular, play has long been a matter of speculation and inquiry. Once regarded as selfish, or parasitic, genomic elements with little or no phenotypic relevance (Doolittle and Sapienza, 1980; Orgel and Crick, 1980), it has since become apparent that TEs make substantial contributions to the structure, function and evolution of their host genomes (Kidwell and Lisch, 2001). Perhaps the most significant functional effect that TEs have had on their host genomes is manifest through the donation of regulatory sequences that control the expression of nearby genes (Feschotte, 2008). Studies of TE regulatory effects have focused, for the most part, on discrete well characterized regulatory elements such as transcription factor binding sites (Jordan et al., 2003; van de Lagemaat et al., 2003; Wang et al., 2007), enhancers (Bejerano et al., 2006) and alternative promoters (Dunn et al., 2003; Conley et al., 2008). A number of recent studies have also outlined the contributions of TEs to regulatory RNA genes (Smalheiser and Torvik, 2005; Borchert et al., 2006; Piriyapongsa and Jordan, 2007; Piriyapongsa et al., 2007). For this study, we sought to analyze the contribution of

Abbreviations: TE, transposable element; SSR, simple sequence repeat; TSS, transcriptional start site; GNF2, Novartis mammalian gene expression atlas 2.

* Corresponding author. Tel.: +1 404 385 2224; fax: +1 404 894 0519.

E-mail address: king.jordan@biology.gatech.edu (I.K. Jordan).

repetitive DNA to epigenetic aspects of gene regulation, specifically the relationship between repetitive DNA elements and the chromatin environment of human promoter sequences.

Genomic DNA in eukaryotes is wrapped around histone proteins and packaged into repeating subunits of chromatin called nucleosomes (Kornberg and Lorch, 1999). The importance of specific genomic sequences in determining the binding locations of nucleosomes has recently been confirmed (Segal et al., 2006). A number of factors point to a relationship between repetitive DNA elements, the local chromatin environment and epigenetic gene regulation. Densely compact heterochromatin is enriched for both TEs and SSRs in a number eukaryotic organisms (Dimitri and Junakovic, 1999). Heterochromatin functions to mitigate potentially deleterious effects associated with TEs by repressing both element transcription and ectopic recombination between dispersed element sequences (Grewal and Jia, 2007). In fact, it has been proposed that heterochromatin originally evolved to serve as a genome defense mechanism by silencing TEs (Henikoff and Matzke, 1997; Henikoff, 2000). In the plant *Arabidopsis*, *de novo* heterochromatin formation can be caused by insertions of TEs into euchromatin, and TEs are able to epigenetically silence genes when they are inserted nearby or inside them (Lippman et al., 2004). In other words, TEs have been shown to cause specific *in situ* changes in the chromatin environment that can spread locally and regulate gene expression in a way that is region-specific but sequence-independent (i.e. epigenetic).

The previously established connections between genome repeats, chromatin environment and gene regulation for model organisms, taken together with the repeat-rich nature of the human genome, suggest that repetitive sequence elements may play a role in regulating human gene expression by modulating the local chromatin environment. Specifically, we hypothesized that gene regulatory related differences in nucleosome binding at human promoter sequences are mediated in part by repetitive genomic elements. We evaluated the relationship between nucleosome binding, repetitive element promoter distributions and human gene expression to test this idea. Human proximal promoter sequences were characterized with respect to both their repetitive DNA architectures and predicted nucleosome binding affinities, and the repetitive DNA environment of the promoters was considered with respect to patterns of gene expression.

2. Materials and methods

2.1. Promoter sequence analysis

Our analysis focused on proximal promoter sequence regions, which we define for a gene as ranging from –1 kb at the 5' end to the transcription start (TSS) at the 3' end. We relied on the Database of Transcriptional Start Sites (DBTSS) to identify experimentally characterized TSS, based on aligned full-length cDNA sequences, in the human genome (Suzuki et al., 2002). These TSS were mapped to the March 2006 human genome reference sequence (NCBI Build 36.1) and used to extract 1 kb proximal promoter sequences as described previously (Marino-Ramirez et al., 2004; Tharakaraman et al., 2005). This procedure was used to ensure analysis of the most accurate set of human proximal promoter sequences possible. For the additional three mammalian species analyzed – chimpanzee (*Pan troglodytes*), mouse (*Mus musculus*) and rat (*Rattus norvegicus*) – the locations of proximal promoter sequences were determined based on the 5' most position of NCBI Refseq gene models (Pruitt et al., 2007). These positions were used to download 1 kb proximal promoter sequences from the latest respective genome builds for each organism from the UCSC Genome Browser (Karolchik et al., 2003): chimpanzee $n=24,170$, mouse $n=20,589$ and rat $n=8737$.

The program RepeatMasker (Smit et al., 1996–2004) was used to detect and annotate repetitive elements in the proximal promoter

sequences. RepeatMasker was run using 500 bp of flanking sequence on either end of the proximal promoter regions analyzed to avoid edge effects in the detection of repeats. Repetitive elements detected by RepeatMasker were broken down into two main categories: interspersed repeats, also known as transposable elements (TEs), and simple sequence repeats (SSRs). SSRs may be annotated as low complexity sequences and correspond to runs of repeating k -mers where $k=1$ –13 bp for microsatellites and $k=14$ –500 for minisatellites. TEs were further divided into specific classes: LINEs, SINEs, LTR and DNA as well as specific families L1 and Alu.

Proximal promoter sequences, including 500 bp flanks, were analyzed using the Nucleosome Prediction software developed by the Segal lab (Segal et al., 2006). This software was used to calculate the probability of each nucleotide being occupied by a nucleosome in all promoter sequences. These nucleosome occupancy probabilities are based on the periodicity of dinucleotides – AA/TT/TA – that are a characteristic of genomic sequences that have been experimentally isolated as bound to nucleosomes. Predictions for the relative placement of nucleosomes along genomic sequence are further informed by a thermodynamic stability model. The nucleosome prediction model used in our analysis is based on experimentally characterized nucleosome bound sequences reported for chicken (Satchwell et al., 1986). The chicken model has been proven accurate when used on other vertebrate genomes (Segal et al., 2006). For sets of promoter sequences, nucleosome occupancy averages were calculated over each position of the 1 kb proximal promoter regions and these average values were taken as the position-specific nucleosome binding affinities (nba) reported here.

Two sets of promoter sequence randomizations were done and position-specific nucleosome binding affinities were re-calculated on the randomized sequence sets. The first randomization consisted of randomly shuffling entire 1 kb proximal promoter sequences. This has the effect of maintaining overall nucleotide composition of the promoter sequences while changing the dinucleotide composition as well as any regional nucleotide biases along the promoters. The second randomization procedure consisted on randomly shuffling non-overlapping 100 bp windows along the promoter sequences in place. This has the effect of maintaining both overall and local nucleotide compositions of the promoters while changing the dinucleotide composition.

2.2. Repeat-based promoter clustering

Human proximal promoter sequences were clustered solely based on their repetitive DNA architectures. To do this, we generated 1000-unit vectors that represent the position-specific repeat content for each promoter sequence. A discrete value was assigned to each promoter sequence position (nucleotide) in the following manner:

$$X_i = \begin{cases} 1 & \text{if the nucleotide is part of a TE sequence} \\ -1 & \text{if the nucleotide is part of a SSR sequence} \\ 0 & \text{if the nucleotide is part of a non-repetitive sequence} \end{cases}$$

where X_i represents the nucleotide at position i .

Promoter sequence repeat vectors were then clustered using a combination of k -means clustering ($k=5, 10, 20$) and Self Organized Mapping using the program Genesis (Sturn et al., 2002). We found that using k -means clustering with $k=5$ followed by a Self Organized Map generated the most coherent clusters in terms of the repeat content of the vectors.

2.3. Gene expression analysis

We used version 2 of the Novartis mammalian gene expression atlas (GNF2), which provides replicate Affymetrix microarray data for 44,775 probes across 79 human tissues (Su et al., 2004). GNF2

expression data, in the form of Affymetrix signal intensity values, were obtained from the UCSC Table Browser (Karolchik et al., 2004), and Affymetrix probes were mapped to NCBI Refseq identifiers using the UCSC Table Browser tools. For each gene, the average, maximum and breadth of expression were computed across the 79 tissues in the GNF2 data set. Expression breadth is taken as the number of tissues where the gene has a signal intensity value of >350. Co-expression between gene pairs was measured by computing the Pearson correlation coefficient (r) between pairs of gene-specific expression signal intensity vectors:

$$g_i = [t_1, t_2 \dots t_{79}]$$

where g_i is the i th gene and t_n is the expression level for that gene in the n th tissue.

For each repeat-specific promoter cluster, the average r -value for all pairwise comparisons between genes in the cluster was computed. In addition, the difference (*diff*) between the cluster-specific r -value averages (cluster- r) and all possible pairwise r -values between genes (all- r) was computed for each cluster:

$$\text{diff} = \text{cluster-}r - \text{all-}r.$$

The significance of these differences was computed using the normal deviate:

$$z = \text{diff} / \text{se}_{\text{diff}}$$

where se_{diff} is the standard error of the difference.

2.4. Probabilistic analysis of promoter repeats

We used a probabilistic representation of the repeat content of the human proximal promoter sequence clusters in order to derive gene (promoter)-specific similarity scores that indicate the probability that any human gene (promoter) belongs to a specific repeat cluster. To do this, each proximal promoter sequence (1 kb upstream of the TSS) in a cluster was divided into 20 non-overlapping windows of 50 bp each. For each window (w), the probability (p) of the occurrence of a TE nucleotide, or SSR nucleotide or a non-repetitive (NR) nucleotide was calculated separately using the following formula:

$$p(b, w) = \frac{f_{b,w} + s(b)}{N + \sum_{b' \in \{T, S, N\}} s(b')}$$

where $f_{b,w}$ = counts of base b in window w and b represents counts of either TE nucleotides, or SSR nucleotides or non-repetitive nucleotides, N = number of sites in the window (50) and $s(b)$ = a pseudocount function. The probabilities thus calculated for each window were averaged for all promoters in the cluster. This procedure was repeated to yield repetitive DNA probabilistic representation models for each of the six promoter clusters.

All the proximal promoter sequences analyzed were then scored against each of the six cluster-specific probabilistic models using a log-likelihood ratio approach illustrated as follows:

$$LL_{b,w} = \ln \sum_{TE, SSR, NR} f_{b,w} \ln \frac{f_{b,w}}{f_b}$$

where $f_{b,w} = p_{b,w} \times 50$, which is the model frequency used as background. Promoter-specific scores (S) were then computed as the sum of log-likelihood ratios over the 20 windows of 50 bp each:

$$S = \sum_{w=1}^{20} LL_{b,w}.$$

Using this method, we scored all genes (promoters) against each of the six cluster models to generate six cluster-specific gene (promoter) score vectors. This modeling and scoring method is a modification of

the approach used to score sequence motifs, such as transcription factor binding sites, based on motif-characteristic position-weight matrices (Wasserman and Sandelin, 2004).

In order to relate promoter sequence repetitive DNA architecture to tissue-specific gene expression, the gene (promoter)-specific probabilistic repeat cluster scores were correlated with tissue-specific gene expression signal intensity values for each of the 79 tissues in GNF. This was repeated with gene (promoter)-specific scores assigned to each gene for each of the six repeat clusters. For example, for the cluster1 ($c1$) versus tissue1 ($t1$) comparison:

$$c1 = [S_{g1}, S_{g2} \dots S_{g7913}] \times t1 = [e_{g1}, e_{g2} \dots e_{g7913}]$$

where g_i is the i th gene, S is the score for the cluster1 model and e is the expression level for that gene in tissue1. In other words, each gene analyzed is assigned a repeat probability score for each of the six clusters, and these six sets of repeat probability promoter scores are individually correlated with the GNF2 tissue-specific expression values for the genes. This procedure resulted in a 6×79 matrix of correlation values.

2.5. Gene Ontology (GO) analysis

GO annotation terms (Ashburner et al., 2000) for human genes were obtained from the Gene Ontology Annotation database (<http://www.ebi.ac.uk/GOA/>). GO terms were further mapped to higher level GO slim categories. Expected versus observed frequencies of GO slim terms were compared using χ^2 tests for each promoter repeat cluster, as well as for the combined TE– and TE+ groups, in order to look for over-represented GO slim categories. The pairwise similarity between GO terms was computed using modified semantic similarity method (Lord et al., 2003; Azuaje et al., 2005) as described previously (Marino-Ramirez et al., 2006; Tsaparas et al., 2006). The GO similarity difference (GO_{diff}) was calculated between the average pairwise similarity for GO terms from pairs of genes within TE groups (e.g. TE +) and the average pairwise GO similarity for all possible pairs of genes:

$$GO_{\text{diff}} = GO_{\text{sim}} - (TE+) - GO_{\text{sim}} - (all).$$

The significance of the difference was measured using the normal deviate as described for the gene expression analysis.

2.6. Statistical analysis

Standard statistical tests were used to compare population means for pairwise (Student's t -test) and for multiple comparisons (ANOVA), to correlated vectors of nucleosome binding affinities, TE and SSR densities, expression and promoter score values (Pearson correlation coefficient), to control for the confounding effects of multiple variables on correlation values obtained (partial correlation) and to evaluate the difference between observed and expected GO terms (χ^2) (Zar, 1999).

3. Results and discussion

3.1. Repetitive DNA and nucleosome binding affinity

Experimentally characterized human gene proximal promoter sequences ($n = 7913$) were taken from the Database of Transcriptional Start Sites (DBTSS) (Suzuki et al., 2002) and analyzed with respect to their repetitive DNA content and nucleosome binding affinities. The locations of repetitive DNA elements along promoter sequences were determined by the RepeatMasker program and nucleosome binding affinities were predicted using the method of (Segal et al., 2006). Two classes of repetitive DNA were analyzed separately: interspersed repeats, also known as transposable elements (TEs) and simple sequence repeats (SSRs), which are made up of runs of exact or nearly

exact repeating *k*-mers. For each promoter position, from 1 kb upstream to the transcriptional start site (TSS), the average TE and SSR densities over all promoter sequences were calculated as the fraction of sequences for which that position was occupied by a TE or SSR. Average nucleosome binding affinities across promoter positions were calculated as the fraction of sequences for which a given position was predicted to be occupied (bound) by a nucleosome. Average nucleosome binding affinities and the average TE density follow parallel trends along the proximal promoter regions (Fig. 1a). Nucleosomes bind more tightly and TEs are found more frequently distal to the TSS, whereas nucleosomes bind promoter sequences most proximal to the TSS with lower affinity and TEs are rarely found close to the TSS. SSRs show a distinctly different trend with a higher density close to the TSS that corresponds to the decrease in nucleosome binding affinity. The SSR density matches the nucleosome binding affinity even more closely than the TE density just upstream of the TSS. Nucleosome binding affinities decrease steadily from distal regions until ~35 bp upstream of the TSS, then the nucleosome binding affinity increases towards the TSS. Similarly, the SSR density increases to the same point and then drops off as the nucleosome binding affinity increases (Fig. 1a). This core promoter region where nucleosome binding affinity is at its lowest and SSR density is at its highest corresponds to the location where the basal transcriptional machinery assembles, and RNA polymerase II binds, to initiate transcription.

The correlations between nucleosome binding affinities with TE and SSR densities along human proximal promoter regions are robust and highly statistically significant (Fig. 1b). Previously, we observed

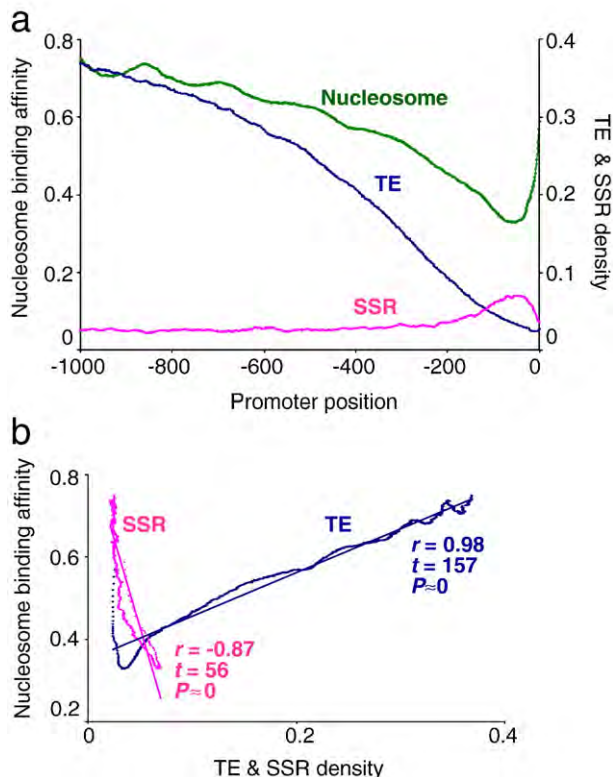


Fig. 1. Repetitive DNA density and nucleosome binding affinity along human proximal promoter sequences. (a) Average nucleosome binding affinities (green line, values on left y-axis) along with average TE densities (blue line, values on right y-axis) and average SSR densities (pink line, values on right y-axis) over 7913 human proximal promoter sequences are plotted over each promoter position starting from -1000 bp upstream and progressing to the transcriptional start site (TSS at position 0). (b) Linear trends and correlations relating position-specific nucleosome binding affinities (y-axis) to TE (blue) and SSR (pink) densities (x-axis) are shown. Statistical significance levels of the *r*-values are based on the Student's *t*-distribution with $df = n - 2 = 998$ where $t = r \cdot \sqrt{(n-2)/(1-r^2)}$.

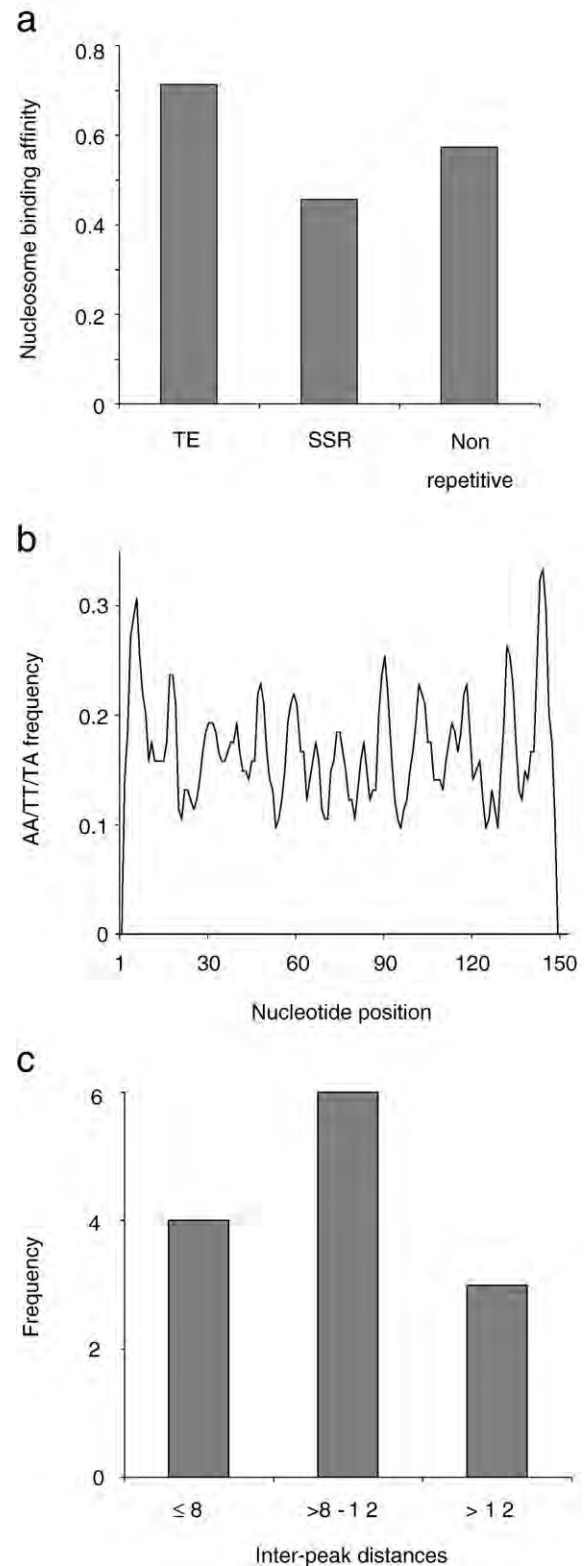


Fig. 2. Nucleosome binding properties for repetitive versus non-repetitive DNA. (a) Average predicted nucleosome binding affinities are shown for TE, SSR and non-repetitive human promoter sequences. (b) Periodicity of the nucleosome binding (wrapping) characteristic dinucleotides AA/TT/TA are shown for 39 experimentally characterized nucleosome bound TE sequences from chicken. (c) Histogram showing the inter-peak distances for AA/TT/TA dinucleotides.

that nucleotide composition changes markedly along human proximal promoter sequences with an increase in CpG frequency close to the TSS (Marino-Ramirez et al., 2004), while the nucleosome binding

Table 1
Average* nucleosome binding affinities for TE classes (families)

TE class (family) ^a	Avg nba \pm s.e. ^b
L1	0.849 \pm 6.8e-4
LINE other	0.805 \pm 7.6e-4
Alu	0.510 \pm 5.2e-4
SINE other	0.789 \pm 7.0e-4
LTR	0.807 \pm 7.9e-4
DNA	0.802 \pm 9.8e-4

^a TEs are broken down by class (family) using RepeatMasker. The L1 and Alu families are considered separately from all other LINEs and SINEs respectively. All LTR and DNA elements are considered together as classes.

^b Average nucleotide binding affinities \pm standard errors.

* All differences are statistically significant (ANOVA, $F=2.8e4$, $P\approx 0$).

prediction method we employed in this analysis relies on the periodicity of AT-rich dinucleotides (Segal et al., 2006). Thus, it is possible that the high (low) nucleosome binding affinity of TE (SSR) sequences in proximal promoter regions is a corollary effect of local differences in nucleotide composition. We attempted to control for this possibility in several ways. First of all, average nucleosome binding affinities were computed for all TE, SSR and non-repetitive sequences irrespective of their locations along proximal promoter regions. On average, TE sequences bind nucleosomes most tightly, followed by non-repetitive DNA and SSRs, which have the lowest nucleosome affinities (Fig. 2a); all differences are highly statistically significant (ANOVA, $F=4.5e11$, $P\approx 0$).

In addition to the binding affinity observations that are based on the nucleosome prediction software, we also analyzed the nucleosome wrapping characteristic AA/TT/TA dinucleotide frequencies along experimentally characterized nucleosome bound sequences from chicken (Satchwell et al., 1986) that we identified as being derived from TEs ($n=39$). The chicken TE sequences show the characteristic AA/TT/TA dinucleotide periodicity expected of nucleosome bound sequences (Fig. 2b); in fact, the average distance between dinucleotide peaks for these TE sequences is ~ 10.3 bp, which is close to the expected distance of 10.2 bp corresponding to one turn of the

DNA helix (Fig. 2c). This is significant because DNA sequences are thought to wrap around nucleosomes by bending sharply at each repeating turn of the DNA helix, and this sharp bending is facilitated by the specific AA/TT/TA dinucleotides (Widom, 2001).

We also attempted to control for nucleotide composition effects by randomizing promoter sequences and re-calculating nucleosome binding affinities. First, entire 1 kb promoter sequences were randomized and nucleosome binding affinities were re-calculated. This control procedure has the effect of eliminating both native dinucleotide occurrences and local nucleotide composition biases. The average nucleotide binding affinity for such randomized promoter sequences ($nba=0.16$) is $\sim 3\times$ lower than seen for the observed promoter sequences ($nba=0.49$), and the difference between random and observed affinities is highly significant ($t=23$, $P=5.3e-100$). In addition to differences in the magnitude of the nucleosome binding affinities, the relative affinity trends along the promoters were compared for the random versus observed sets. Partial correlation was used to control for the effects of the random sequences on the observed relationship between nucleosome binding affinity with TE and SSR densities along proximal promoters. The positive (negative) correlations between nucleosome binding for TE (SSR) do not change when the correlations between random sequences and nucleosome binding along the promoters are accounted for [$r_{nba-TE|random1}=0.99$ and $r_{nba-SSR|random1}=0.85$].

A second randomization procedure was done to account for local differences in nucleotide composition along proximal promoter sequences. In this case, sequences were randomized within non-overlapping 100 bp windows along the promoters. This had the effect of eliminating native dinucleotide occurrences while maintaining local nucleotide composition. As with the complete sequence randomization procedure, the locally randomized sequences have significantly lower nucleosome binding affinities ($nba=0.23$) than the observed sequences ($nba=0.49$), and this $2.1\times$ difference is highly statistically significant ($t=17$, $P=5.0e-55$). Clearly, local nucleotide composition alone cannot explain the observed nucleosome binding affinities. However, the relative trends in nucleosome binding show different

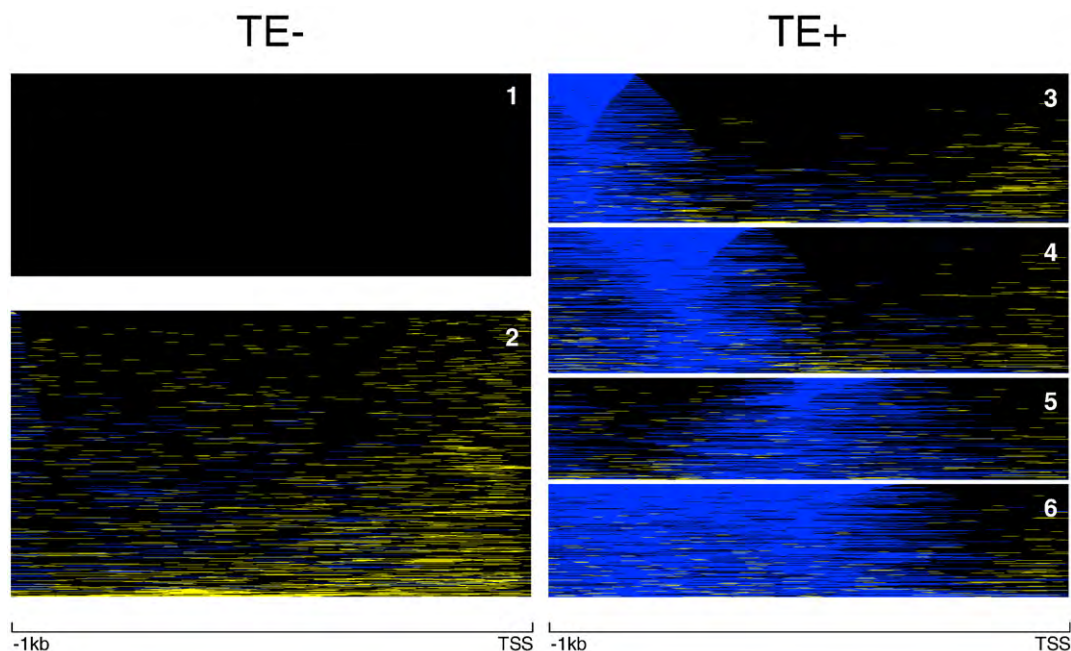


Fig. 3. Clusters of human proximal promoters based on their repetitive DNA sequence distributions. Proximal promoter sequences are represented left-to-right from position -1000 bp upstream to the transcriptional start site (TSS). Promoter sequences are color coded according to their repeat element distributions. Individual promoter nucleotide positions occupied by TEs are shown in blue, SSR positions are shown in yellow and non-repetitive positions are shown in black. The vertical size of the clusters corresponds to the number of sequences in each cluster. There are two (c1 and c2) clusters that contain promoters largely devoid of TE sequences (TE-), and the promoter sequences of the remaining four clusters (TE+, c3–c6) contain increasing numbers of TEs.

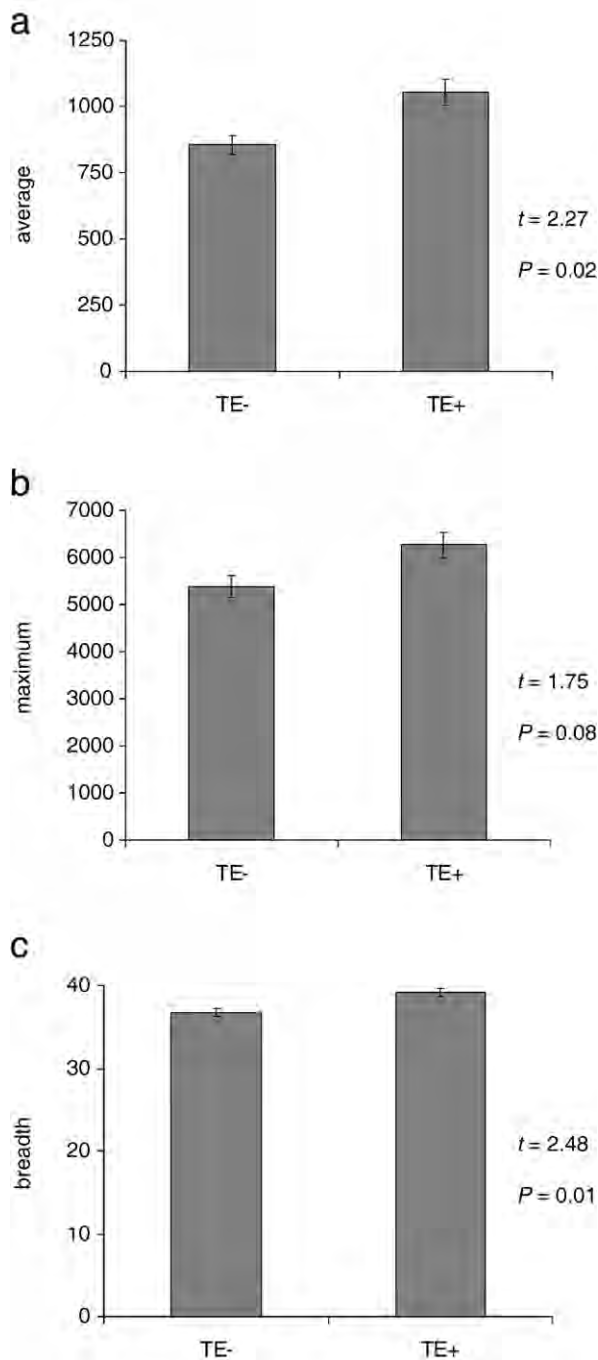


Fig. 4. Gene expression comparison for TE- versus TE+ promoter clusters. Human gene expression data are from the Novartis mammalian gene expression atlas version 2 (GNF2). (a) Average level of expression, (b) maximum level of expression and (c) breadth of expression across 79 human tissues (cells) are compared for genes that have TE- versus TE+ promoter sequences. Statistical significance levels are based on the Student's *t*-test.

local nucleotide composition effects for TEs versus SSRs. The partial correlation controlling for the effects of local nucleotide composition on the relationship between TE density and nucleosome binding eliminates the positive correlation seen across the entire promoter for the observed data [$r_{\text{nba} \cdot \text{TE} | \text{random2}} = -0.14$]. This suggests that local nucleotide composition bias influences the decreasing trend in nucleosome binding affinities along proximal promoters irrespective of TE density. Interestingly, this same mitigating effect of local nucleotide composition is not seen for the relationship between SSRs and nucleosome binding [$r_{\text{nba} \cdot \text{SSR} | \text{random2}} = -0.53$]. This suggested the

possibility that most of the local nucleotide composition bias effect on the relationship between TEs and nucleosome binding may be confined to the region closest to the TSS where TEs are largely absent and SSRs are at their most dense (Fig. 1a). Indeed, when partial correlation controlling for local nucleotide bias is done excluding 150 bp upstream of the TSS, the positive correlation between TEs and nucleosome binding affinity remains [-1000 to -150 $r_{\text{nba} \cdot \text{TE} | \text{random2}} = 0.76$]. In other words, positive TE effects on nucleosome binding are most evident away from the TSS, while the SSRs that inhibit nucleosome binding act closest to the TSS.

Taken together, these data suggest the intriguing possibility that the human genome utilizes repetitive DNA content along promoter regions to tune nucleosome binding in such a way as to facilitate maximum access of the basal transcriptional machinery just upstream of TSS. Furthermore, different classes of repeats play distinct roles in this process; TEs bind nucleosomes tightly yielding compact less accessible DNA, while SSRs extrude nucleosomes creating a relatively open chromatin environment.

3.2. Cross-species comparison

In addition to the human genome analysis, the relationship between nucleosome binding and repetitive DNA content of proximal promoter regions was evaluated for four additional mammalian species with complete genome sequences available: chimpanzee (*P. troglodytes*), mouse (*M. musculus*) and rat (*R. norvegicus*). For these species, NCBI Refseq gene models were used to define TSS and proximal promoter regions, while TE and SSR repeats and nucleosome binding were analyzed as was done for the human genome. The trends observed for human are highly similar to those seen for the other mammalian species (Supplementary Fig. 1). In chimpanzee, mouse and rat, nucleosome binding affinities decrease steadily along the proximal promoter region until the core promoter, <50 bp from the TSS, where nucleosome binding begins to increase. For these three species, TE density drops precipitously and steadily along the proximal promoter while SSR density increases sharply at first in the core promoter near the TSS and then drops off again as nucleosome binding affinity increases. Thus, repeat-rich mammalian genomes appear to use repetitive DNA elements to tune nucleosome binding and core promoter accessibility in similar ways. The conservation of the relationship between repetitive DNA and nucleosome binding in core promoters of several mammalian species suggests that this mechanism may have evolved early in the mammalian radiation as repetitive

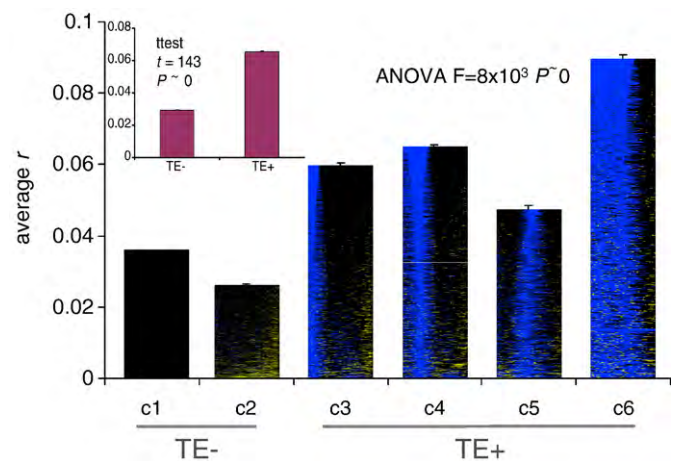


Fig. 5. Gene co-expression for repeat-specific proximal promoter clusters. Average pairwise Pearson correlation coefficients (*r*) for gene expression across 79 human tissues are shown for clusters 1–6 (see Fig. 3) as well as for the TE- versus TE+ clusters (inset). Statistical significance levels are based on ANOVA for multiple comparisons and on the Student's *t*-test for the TE- versus TE+ comparison.

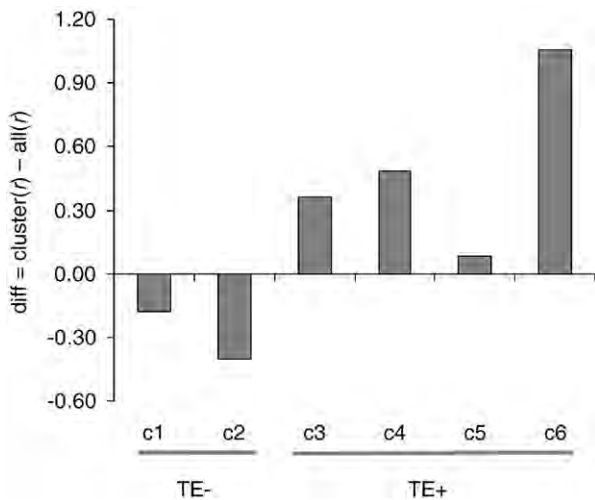


Fig. 6. Differences in gene co-expression between cluster-specific gene pairs versus all possible pairs of genes. Average pairwise Pearson correlations (r) for gene expression across 79 human tissues were measured for all possible gene pairs and this value was subtracted from the average pairwise r -values for genes within each repeat-specific cluster (c1–c6). A negative value indicates that genes within the cluster have less similar co-expression than background, whereas a positive value indicates that genes within a cluster are more highly co-expressed than expected.

elements were proliferating within genomes. However, many of the repetitive elements that yield these patterns evolve rapidly and are lineage-specific. Accordingly, there may be an ongoing dynamic between repeat generation by mutation and/or transposition followed by selection based on the promoter location of the repeat and specific requirements for chromatin accessibility. For TEs in particular, this could simply mean that the elements are eliminated from core promoter regions close to the TSS by purifying selection. Indeed, negative selection against TE insertions closest to TSS would seem to be the easiest way to explain the observed pattern of TE density (Fig. 1a and Supplementary Fig. 1). However, our analysis of gene expression data, described in following sections, suggests that this is not the case. SSRs, on the other hand, appear to be favored in core promoter regions.

3.3. TE-specific effects on nucleosome binding affinity

The Repbase library of repetitive DNA elements used by the program RepeatMasker can be used to annotate TEs into different classes and families (Jurka et al., 2005; Kapitonov and Jurka, 2008). Using this approach, human TE sequences were divided into LINES, (L1 and other LINES), SINEs (Alu and other SINEs), LTR retrotransposons, and DNA transposons to determine if different classes (families) of elements show differential nucleosome binding affinities (Table 1). In general, LINES, LTR retrotransposons and DNA transposons have higher affinities for nucleosomes compared to SINEs. Specifically, L1 elements exhibit the highest nucleosome binding affinities while Alu elements display the lowest affinity for nucleosomes. All differences are statistically significant (Table 1, ANOVA).

The differences in nucleosome binding affinities between L1 and Alu are consistent with their respective nucleotide compositions and perhaps also relevant to their genomic distributions. L1 elements, and LINES in general, are more AT-rich than Alus (SINEs), and AT-rich sequences are more likely to bind nucleosomes tightly as discussed previously. L1 elements are also biased towards intergenic regions in their distribution, while Alu elements are found primarily in gene rich regions. In fact, it has been shown that Alus are preferentially retained in GC- and gene-rich regions of the genome, and this has been taken to suggest that they may be selectively fixed therein by virtue of some gene-related function that they play (Lander et al., 2001). Our data showing lower nucleosome binding for Alu elements suggests that they may be retained in gene regions by virtue of their ability to maintain a relatively open chromatin environment. Conversely, L1 elements may help to maintain compact chromatin structure characteristic of intergenic regions.

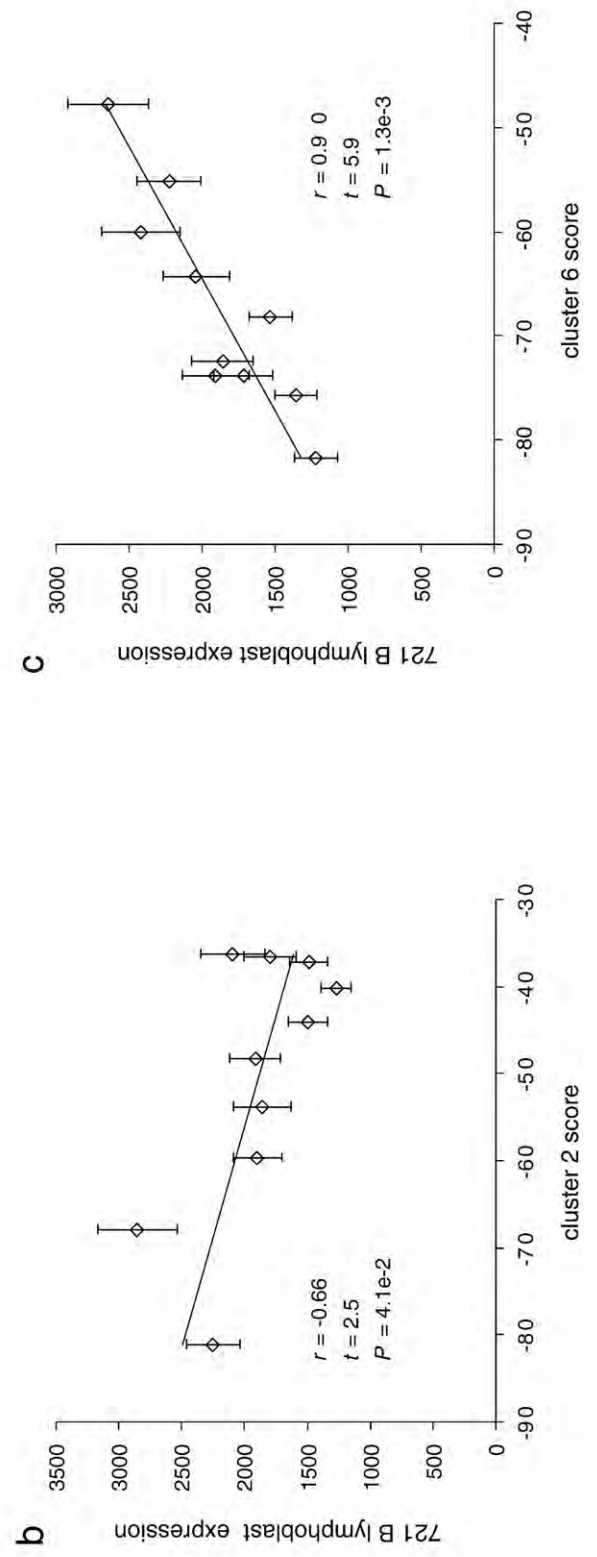
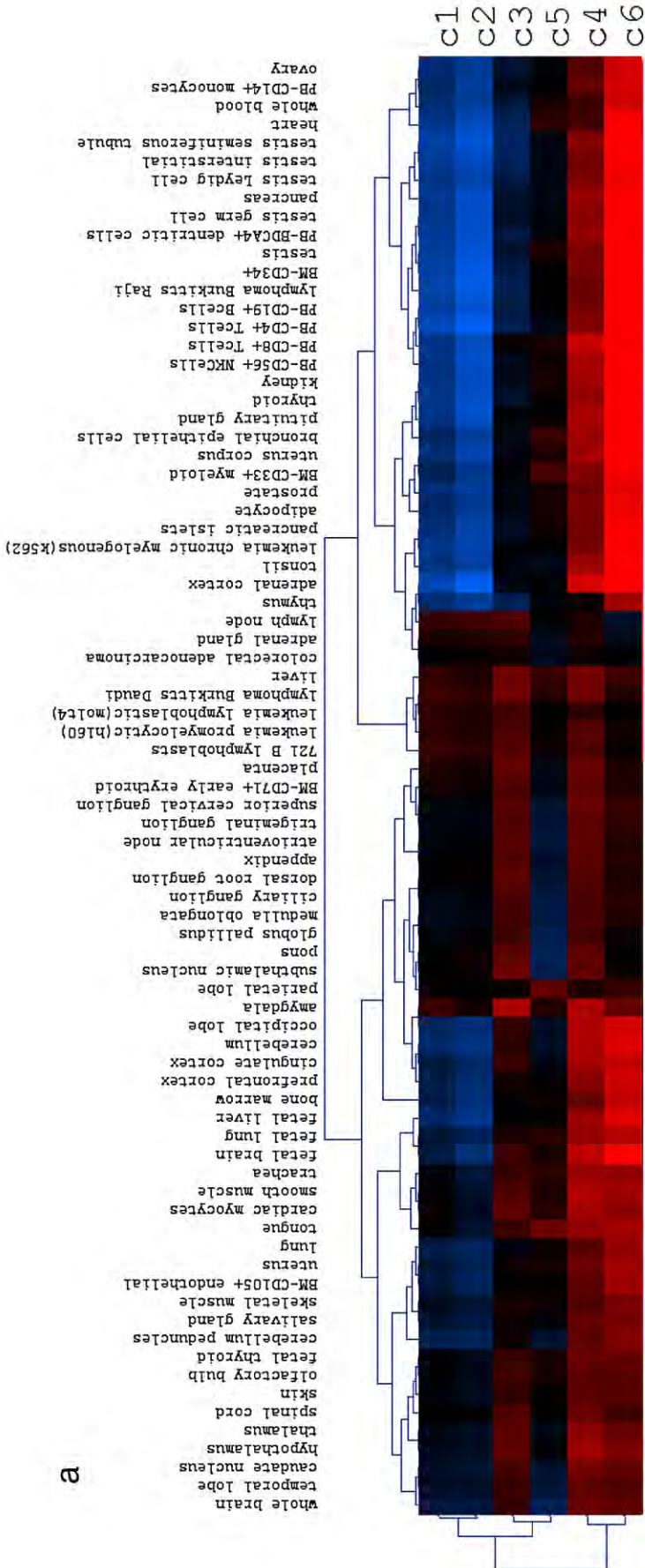
3.4. Promoter repeat architecture and gene expression levels

In light of the observed relationship between repetitive DNA elements and nucleosome binding, we used the repetitive DNA content of proximal promoter regions to group human genes into related clusters. The gene expression and functional properties of the clusters were then compared to their characteristic repeat architectures. To cluster human genes using their promoter repeat distributions, proximal promoter sequences were represented as 1000-unit vectors with each position in a sequence-specific vector receiving a score indicating whether that particular nucleotide is part of a TE, SSR or non-repetitive sequence. These gene-specific promoter repeat vectors were then compared using a distance metric and clustered as described (Materials and methods). This approach ensured that the clusters reflect both the abundance, or lack thereof, and the location of distinct repetitive DNA elements in human promoter sequences. In other words, this scheme relates human genes solely by virtue of their promoter repeat distributions.

We obtained six repeat-specific clusters of human genes in this way (Fig. 3), each cluster representing a distinct overall pattern of TE and/or SSR content and distribution. Two of these clusters (c1 and c2, TE-) consist of genes that are largely devoid of TEs, while four consist of genes with increasing TE densities (c3–c6, TE+). c1 does not contain any repetitive DNA, while c2 is enriched in SSR sequences and has very low TE content. c3–c6 have progressively more TE content with locations shifting slightly towards the TSS.

The gene expression properties of the human genes in these clusters were analyzed using version 2 of the Novartis mammalian gene expression atlas (GNF2) (Su et al., 2004). This data set consists of Affymetrix microarray experiments, performed in replicate, on 79 different human tissue (cell) samples. For each human gene, over 79 tissues, we computed the average expression level, maximum expression level and breadth of expression as described (Materials and methods); cluster-specific averages for each of these parameters were then compared (Fig. 4). We were surprised to find that clusters that contain TEs (c3–c6, TE+) have higher average, maximum and breadth of expression than clusters that are largely devoid of TEs (c1 and c2, TE-). Gene expression levels are known to correlate with a

Fig. 7. Promoter repetitive DNA architecture and tissue-specific gene expression. Probabilistic models were used to represent the repetitive DNA architectures of each repeat-specific cluster (see Fig. 3 and Supplementary Fig. 2). Cluster-specific probabilistic models were used to score individual promoter sequences in terms of how closely they resemble a given cluster (Materials and methods). Vectors of cluster-specific gene scores were correlated with vectors of gene expression values specific human tissues. (a) A heat map illustrating the relative correlation values between gene (promoter)-specific scores for each cluster and tissue-specific gene expression values for the 79 tissues in the Novartis gene expression atlas version 2 (GNF2). Relatively high (positive) correlations between gene-cluster scores and gene expression levels are shown in red and low (negative) correlations are shown in blue. Two specific examples of such correlations are shown in panels b and c. (b) Gene (promoter)-specific scores based on the probabilistic model for cluster 2 are negatively correlated with gene expression levels in a B lymphoblast cell line. (c) Gene (promoter)-specific scores based on the probabilistic model for cluster 6 are positively correlated with gene expression levels in a B lymphoblast cell line. In other words, genes with repetitive DNA promoter profiles that most closely resemble cluster 6 are more highly expressed in the B lymphoblast cell line, whereas genes with repetitive DNA promoter profiles that resemble cluster 2 have lower levels of B lymphoblast expression.



number of measures of gene ‘importance’ such as sequence and phylogenetic conservation, fitness effects, numbers of protein interactions, etc. (Duret and Mouchiroud, 2000; Pal et al., 2001; Krylov et al., 2003; Zhang and Li, 2004; Wolf et al., 2006). In other words, genes that are more highly and broadly expressed are under greater purifying selection than genes with lower expression levels. If TEs are eliminated from proximal promoter sequences by purifying selection, then one may expect that TE+ promoters would have lower, and not higher as we observe, levels of gene expression than TE– promoters. In other words, our analysis of repeat cluster gene expression levels argues against the straightforward interpretation that the paucity of TEs in proximal promoter sequences, and their decreasing frequency closer to TSS, is a result of purifying selection against disruptive insertions in core promoters.

On the other hand, one may expect that genes with more restricted and more tightly regulated expression, such as developmental genes, would have more TE sensitive promoters than genes that are highly and broadly expressed. In fact, developmental genes are known to have promoters that are largely devoid of TEs (Simons et al., 2006, 2007). This may reflect the fact that such genes are more finely and tightly regulated and accordingly contain more complex promoters with higher numbers of cis-regulatory elements. If this is indeed the case, then the paucity of TEs in proximal promoter regions may still be explained, to some extent, by purifying selection against disruptive insertions. Discrimination between these two hypotheses regarding the selective elimination, or lack thereof, of proximal promoter TE sequences awaits further analysis.

3.5. Promoter repeat architecture and tissue-specific gene co-expression

In addition to analyzing repeat cluster gene expression levels, we also evaluated the relationship between the tissue-specific expression patterns of genes across the 79 tissues from GNF2 and their promoter repeat content. To do this, gene-specific vectors of expression levels across tissues were compared using the Pearson correlation coefficient (r); positive values of r indicate gene pairs that are co-expressed across tissues. For each cluster, average r -values were computed based on all pairwise comparisons within the cluster (Fig. 5). Higher average r -values are associated with increasing TE promoter content of the clusters. For instance, there is a positive ($R = 0.77$), albeit marginally significant ($z = 1.72$, $P = 0.1$), rank correlation between cluster TE content and co-expression. In addition, all four TE+ clusters have greater average co-expression than either of the TE– clusters, and the average r -value for TE+ clusters together is significantly greater than seen for the combined TE– clusters (Fig. 5).

The possibility of gene co-regulation within repeat clusters was also evaluated by taking the difference between the average r -value for all pairwise comparisons within clusters to average pairwise r -value for all gene comparisons (Materials and methods) (Fig. 6). If genes within clusters are co-regulated, then the value of this difference should be positive, whereas no co-regulation will yield a negative difference value. The TE– clusters 1 and 2 have negative difference values indicating that genes with no TEs in their promoters are less co-expressed with other genes possessing a similar lack of repeats than they are with all genes. On the other hand, the TE+ clusters 3–6 all have positive difference values further demonstrating that genes with similar repetitive DNA profiles in their promoters are more closely co-expressed than random pairs of genes. The difference values for each cluster are statistically significant ($7.3 < z < 100.6$, $1.4e-13 < P < 0$).

Taken together, these observations on gene co-expression also argue against the notion that TE insertions in proximal promoter sequences are basically disruptive or deleterious, since the presence of similar TE promoter distributions implies a higher level of gene co-regulation than the absence of TEs does. This is not to say that the majority of *de novo* TE

insertions in and around functional promoter sequences are not deleterious, clearly they are. However, the repeat sequences that have been fixed in proximal promoter sequences do appear to make functionally relevant contributions to chromatin accessibility and help to regulate levels and specific patterns of gene expression.

3.6. Probabilistic analysis of promoters and gene expression

Given the relationship between gene expression and the repetitive DNA architecture of human promoters we observed, we wanted to further evaluate the propensity of human genes to be expressed in specific tissues based on the repetitive DNA content of their promoters. To do this, we used a probabilistic representation of cluster-specific promoter architectures together with the GNF2 expression data. This involved partitioning 1 kb proximal promoter sequences into 20 non-overlapping windows of 50 bp each, and for a given cluster, representing the probability of observing TE, SSR or non-repetitive nucleotides in each window (Materials and methods). The probabilistic representation of promoter repeat architectures we employed is mathematically analogous to the probabilistic representations of position weight matrices (PWMs) used to summarize position-specific residue frequencies among collections of sequence motifs such as transcription factor binding sites (Wasserman and Sandelin, 2004). Accordingly, promoter repeat profiles can be represented as sequence logos showing the probability and distribution for sites of different repeat classes (Supplementary Fig. 2). The cluster-specific promoter repeat profiles can then be used to score individual promoter sequences just as PWM representations can be used to score putative motif sequences. Connecting these cluster- and position-specific promoter repeat profiles to tissue-specific gene expression profiles was done in a way that is similar to the methodology used to connect the presence of transcription factor binding site motifs to specific gene expression patterns (Conlon et al., 2003).

For each of the 79 tissues in GNF2, each promoter sequence was given six cluster-specific scores, and for each cluster, the gene-specific scores were correlated with the tissue-specific gene expression levels (Materials and methods). This resulted in a 6-by-79 matrix of cluster-by-tissue correlations (Fig. 7). The TE+ clusters 4 and 6 show particularly high correlations with a number of tissues, such as B lymphoblasts (Figs. 7b and c), whereas the TE– clusters 1 and 2 show low correlations with the same tissues and lower correlations overall. This indicates that certain repeat-rich promoter architectures play a role in driving tissue-specific expression, while repeat poor promoters have less coherent regulatory properties. In addition, the differences in promoter score-expression level correlations across tissues and between clusters indicate that different repeat contexts are likely to have tissue-specific regulatory functions. Hierarchical clustering of the tissues and the clusters, according to the promoter score-expression level correlations, group related tissues together including reproductive tissues, immune related cells and cancer samples (Fig. 7a). This indicates that TE-rich promoters may help to regulate genes that function specifically in these tissues further underscoring the biological significance of promoter sequence repetitive DNA profiles.

3.7. Gene Ontology analysis

Having established a connection between repetitive DNA promoter architectures and gene regulation, we wondered whether genes with similar promoter repeat distributions encoded proteins with related functions. In order to test this, we used analysis of Gene Ontology (GO) terms for genes within and between the TE– versus the TE+ repeat-specific promoter clusters (Fig. 3). A modified version of the GO semantic similarity measure (Lord et al., 2003; Azuaje et al., 2005) was used to compare the similarities between GO terms within clusters versus the background GO similarity among all pairs of genes. As described previously (Marino-Ramirez et al., 2006; Tsaparas et al.,

Table 2Over-represented* GO slim^a terms for repeat-specific promoter clusters

Group ^b	Molecular function ^c	Cellular component ^c	Biological process ^c
TE–	GO:0030528: transcription regulator activity	–	GO:0007154: cell communication GO:0007275: multicellular organismal development GO:0050789: regulation of biological process GO:0006810: transport GO:0007154: cell communication
TE+	GO:0003824: catalytic activity GO:0016491: oxidoreductase activity	GO:0005737: cytoplasm	GO:0007154: cell communication GO:0007275: multicellular organismal development GO:0007610: behavior GO:0030154: cell differentiation GO:0050789: regulation of biological process
C1	GO:0005198: structural molecule activity	–	–
C2	GO:0016301: kinase activity GO:0016491: oxidoreductase activity GO:0030528: transcription regulator activity	–	GO:0007154: cell communication GO:0007275: multicellular organismal development GO:0007610: behavior GO:0030154: cell differentiation GO:0050789: regulation of biological process
C3	–	–	–
C4	GO:0003824: catalytic activity	GO:0005737: cytoplasm	GO:0006944: membrane fusion GO:0009056: catabolic process GO:0050896: response to stimulus
C5	GO:0004872: receptor activity GO:0005215: transporter activity GO:0022857: transmembrane transporter activity	GO:0009986: cell surface	GO:0008152: metabolic process GO:0009058: biosynthetic process
C6	GO:0003824: catalytic activity	GO:0005622: intracellular GO:0005737: cytoplasm	GO:0008152: metabolic process GO:0009058: biosynthetic process

^a GO slim categories provide a high level view of GO functions and subsume a number of lower (more granular) GO functional annotation categories.^b Repeat-specific clusters 1–6 along with the combined TE+ and TE– groups (see Fig. 3).^c GO functional annotation categories.* Statistical significance for over-represented terms was evaluated using with χ^2 tests with at least $\chi^2 > 4.2$, $P < 0.04$.

2006), the GO semantic similarity approach measures the pairwise similarity between annotation terms along the GO directed acyclic graph in order to evaluate the functional similarity between pairs of genes. For TE– and TE+ genes, the GO similarity difference (*GOdiff*) is equal to the average GO similarity for all gene pairs within clusters minus the average GO similarity for all possible gene pairs (Materials and methods). Negative values of *GOdiff* indicate that gene pairs are more similar within clusters than for all possible pairs. Both the TE– and TE+ gene sets encode proteins that are significantly more functionally similar than the background comparison set [TE– = $-3.4e-3$, $z = 34$, $P \approx 0$; TE+ = $-7.9e-3$, $z = 11$, $P = 4.8e-3$]. However, within the TE+ clusters, pairs of genes encode proteins that are significantly more functionally similar, on average, than the pairs of genes found within the TE– clusters ($t = 5.8$, $P = 6.4e-9$). This is consistent with the stronger signal of gene co-regulation seen for clusters of promoter sequences that are enriched for TEs and underscores the potential biological significance of repeat-rich promoter sequences in the human genome.

Given the functional coherence of repeat-specific clusters demonstrated by the GO similarity analysis, we wanted to evaluate whether certain GO functional categories are over-represented within specific clusters. To do this, we traced the GO terms represented in the dataset to GO slim terms (Table 2). GO slim categories provide a higher level view of more granular individual GO annotations in order to provide an overview of the kinds of functions that may be over-represented in different groups. The observed counts of GO slim categories for each of the six repeat-specific clusters, as well as for the combined TE– and TE+, groups were compared to their expected values based on the background GO slim frequencies across all clusters to look for over-represented terms. Genes in the electron transport, cytoplasm, catalytic activity and oxidoreductase activity categories were found to be over-represented in TE+ clusters and accordingly under-represented in the TE– clusters, whereas genes in cell communication, multicellular organismal development, regulation of biological process and transcription regulator activity categories are over-represented in TE– clusters and under represented in TE+ clusters. Evaluation of over-represented GO terms in individual clusters reveals coherence across the three categories of GO terms: molecular function, cellular component and biological process. For instance, the TE+ cluster 5 has an over-represented receptor and transporter activities in the molecular function category that agree with the cell surface cellular component term and the response to stimulus biological process term. The over-

represented catalytic activity molecular process term for the most TE-rich cluster 6 corresponds to a cytoplasmic cellular component term along with metabolic and biosynthetic biological process terms. In a general sense, the coherence of GO functional annotations within repeat-specific clusters and the differences between clusters are consistent with biological significance of the regulatory differences seen for these clusters.

4. Conclusion

We have uncovered a connection between repetitive DNA sequences and nucleosome binding in human proximal promoter regions along with an influence of repetitive DNA promoter sequences on specific patterns of gene expression. Interestingly, different classes of repetitive elements function differently to mediate nucleosome binding; TEs bind nucleosomes tightly and are generally excluded from core promoter regions, while SSRs have a low affinity for nucleosomes and are enriched just upstream of TSSs. Thus, it appears that repetitive sequence elements are differentially utilized to tune the accessibility to promoter sequences by transcription factors, particularly the basal transcriptional machinery that assembles just upstream of the TSS, via changes in the local chromatin environment.

Acknowledgement

This research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI. LMR is supported by Corporacion Colombiana de Investigacion Agropecuaria - CORPOICA. IKJ was supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). The authors would like to thank Lee S. Katz and Jittima Piriyaopongsa for helpful discussions and technical advice.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2009.01.013.

References

Ashburner, M., et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

- Azuaje, F., Wang, H., Bodenreider, O., 2005. Ontology-driven similarity approaches to supporting gene functional assessment. *Proc ISMB SIG meeting on Bio-ontologies* 2005, 9–10.
- Bejerano, G., et al., 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441, 87–90.
- Borchert, G.M., Lanier, W., Davidson, B.L., 2006. RNA polymerase III transcribes human microRNAs. *Nat. Struct. Mol. Biol.* 13, 1097–1101.
- Britten, R.J., Kohne, D.E., 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* 161, 529–540.
- Conley, A.B., Piriyaopongsa, J., Jordan, I.K., 2008. Retroviral promoters in the human genome. *Bioinformatics* 24, 1563–1567.
- Conlon, E.M., Liu, X.S., Lieb, J.D., Liu, J.S., 2003. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3339–3344.
- Dimitri, P., Junakovic, N., 1999. Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. *Trends Genet.* 15, 123–124.
- Doolittle, W.F., Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284, 601–603.
- Dunn, C.A., Medstrand, P., Mager, D.L., 2003. An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12841–12846.
- Duret, L., Mouchiroud, D., 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* 17, 68–74.
- Feschotte, C., 2008. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* 9, 397–405.
- Grewal, S.I., Jia, S., 2007. Heterochromatin revisited. *Nat. Rev. Genet.* 8, 35–46.
- Henikoff, S., 2000. Heterochromatin function in complex genomes. *Biochim. Biophys. Acta* 1470, O1–8.
- Henikoff, S., Matzke, M.A., 1997. Exploring and explaining epigenetic effects. *Trends Genet.* 13, 293–295.
- Jordan, I.K., Rogozin, I.B., Glazko, G.V., Koonin, E.V., 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* 19, 68–72.
- Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467.
- Kapitonov, V.V., Jurka, J., 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* 9, 411–412 author reply 414.
- Karolchik, D., et al., 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* 31, 51–54.
- Karolchik, D., et al., 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32, D493–D496.
- Kidd, J.M., et al., 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* 453, 56–64.
- Kidwell, M.G., Lisch, D.R., 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int. J. Org. Evolution* 55, 1–24.
- Kornberg, R.D., Lorch, Y., 1999. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 98, 285–294.
- Krylov, D.M., Wolf, Y.I., Rogozin, I.B., Koonin, E.V., 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* 13, 2229–2235.
- Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Lippman, Z., et al., 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430, 471–476.
- Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A., 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19, 1275–1283.
- Marino-Ramirez, L., Spouge, J.L., Kanga, G.C., Landsman, D., 2004. Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res.* 32, 949–958.
- Marino-Ramirez, L., Bodenreider, O., Kantz, N., Jordan, I.K., 2006. Co-evolutionary rates of functionally related yeast genes. *Evol. Bioinform Online* 2, 295–300.
- Nishihara, H., Smit, A.F., Okada, N., 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res.* 16, 864–874.
- Orgel, L.E., Crick, F.H., 1980. Selfish DNA: the ultimate parasite. *Nature* 284, 604–607.
- Pal, C., Papp, B., Hurst, L.D., 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158, 927–931.
- Piriyaopongsa, J., Jordan, I.K., 2007. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE* 2, e203.
- Piriyaopongsa, J., Marino-Ramirez, L., Jordan, I.K., 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* 176, 1323–1337.
- Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, D61–D65.
- Satchwell, S.C., Drew, H.R., Travers, A.A., 1986. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* 191, 659–675.
- Segal, E., et al., 2006. A genomic code for nucleosome positioning. *Nature* 442, 772–778.
- Simons, C., Pheasant, M., Makunin, I.V., Mattick, J.S., 2006. Transposon-free regions in mammalian genomes. *Genome Res.* 16, 164–172.
- Simons, C., Makunin, I.V., Pheasant, M., Mattick, J.S., 2007. Maintenance of transposon-free regions throughout vertebrate evolution. *BMC Genomics* 8, 470.
- Smalheiser, N.R., Torvik, V.I., 2005. Mammalian microRNAs derived from genomic repeats. *Trends Genet.* 21, 322–326.
- Smit, A.F.A., Hubley, R. and Green, P. (1996–2004), RepeatMasker. p. <http://repeatmasker.org>.
- Sturn, A., Quackenbush, J., Trajanoski, Z., 2002. Genesis: cluster analysis of microarray data. *Bioinformatics* 18, 207–208.
- Su, A.I., et al., 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6062–6067.
- Suzuki, Y., Yamashita, R., Nakai, K., Sugano, S., 2002. DBTSS: database of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.* 30, 328–331.
- Tharakaraman, K., Marino-Ramirez, L., Sheetlin, S., Landsman, D., Spouge, J.L., 2005. Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics* 21 (Suppl. 1) i440–8.
- Tsapas, P., Marino-Ramirez, L., Bodenreider, O., Koonin, E.V., Jordan, I.K., 2006. Global similarity and local divergence in human and mouse gene co-expression networks. *BMC Evol. Biol.* 6, 70.
- van de Lagemaat, L.N., Landry, J.R., Mager, D.L., Medstrand, P., 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* 19, 530–536.
- Venter, J.C., et al., 2001. The sequence of the human genome. *Science* 291, 1304–1351.
- Wang, H., et al., 2005. SVA elements: a hominid-specific retroposon family. *J. Mol. Biol.* 354, 994–1007.
- Wang, T., et al., 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci. U. S. A.* 104, 18613–18618.
- Wasserman, W.W., Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 5, 276–287.
- Widom, J., 2001. Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.* 34, 269–324.
- Wolf, Y.I., Carmel, L., Koonin, E.V., 2006. Unifying measures of gene function and evolution. *Proc. Biol. Sci.* 273, 1507–1515.
- Zar, J.H., 1999. *Biostatistical Analysis*, Fourth ed. Prentice-Hall, Upper Saddle River.
- Zhang, L., Li, W.H., 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.* 21, 236–239.

Chapter 1

Identification of *cis*-Regulatory Elements in Gene Co-expression Networks Using A-GLAM

Leonardo Mariño-Ramírez, Kannan Tharakaraman, Olivier Bodenreider, John Spouge, and David Landsman

Abstract

Reliable identification and assignment of *cis*-regulatory elements in promoter regions is a challenging problem in biology. The sophistication of transcriptional regulation in higher eukaryotes, particularly in metazoans, could be an important factor contributing to their organismal complexity. Here we present an integrated approach where networks of co-expressed genes are combined with gene ontology-derived functional networks to discover clusters of genes that share both similar expression patterns and functions. Regulatory elements are identified in the promoter regions of these gene clusters using a Gibbs sampling algorithm implemented in the A-GLAM software package. Using this approach, we analyze the cell-cycle co-expression network of the yeast *Saccharomyces cerevisiae*, showing that this approach correctly identifies *cis*-regulatory elements present in clusters of co-expressed genes.

Key words: Promoter sequences, transcription factor-binding sites, co-expression, networks, gene ontology, Gibbs sampling.

1. Introduction

The identification and classification of the entire collection of transcription factor-binding sites (TFBSs) are among the greatest challenges in systems biology. Recently, large-scale efforts involving genome mapping and identification of TFBS in lower eukaryotes, such as the yeast *Saccharomyces cerevisiae*, have been successful (1). On the other hand, similar efforts in vertebrates have proven difficult due to the presence of repetitive elements and an increased regulatory complexity (2–4). The accurate prediction and identification of regulatory elements in higher eukaryotes remains a challenge for computational biology, despite recent

progress in the development of algorithms for this purpose (5). Typically, computational methods for identifying *cis*-regulatory elements in promoter sequences fall into two classes, enumerative and alignment techniques (6). We have developed algorithms that use enumerative approaches to identify *cis*-regulatory elements statistically significantly over-represented in promoter regions (7). Subsequently, we developed an algorithm that combines both enumeration and alignment techniques to identify statistically significant *cis*-regulatory elements positionally clustered relative to a specific genomic landmark (8).

Here, we will present a systems biology framework to study *cis*-regulatory elements in networks of co-expressed genes. This approach includes a network comparison operation, namely the intersection between co-expression and functional networks to reduce complexity and false positives due to co-expression linkage but absence of functional linkage. First, co-expression (9, 10) and functional networks (11, 12) are created using user-selected thresholds. Second, the construction of a single network is obtained from the intersection between co-expression and functional networks (13). Third, the highly interconnected regions in the intersection network are identified (14). Fourth, upstream regions of the gene clusters that are linked by both co-expression and function are extracted. Fifth, candidate *cis*-regulatory elements using A-GLAM (8) present in dense cluster regions of the intersection network are identified. In principle, the calculation of intersections for other types of networks with co-expression and/or functional networks could also be used to identify groups of co-regulated genes of interest (15) that may share *cis*-regulatory elements.

2. Materials

2.1. Hardware Requirements

1. Personal computer with at least 512 MB of random access memory (RAM) connected to the Internet.
2. Access to a Linux or UNIX workstation.

2.2. Software Requirements

1. The latest version of the Java Runtime Environment (JRE) freely available at <http://www.java.com/>.
2. The latest version of Cytoscape – a bioinformatics software platform for visualizing molecular interaction networks (13) freely available at <http://www.cytoscape.org/>.
3. The latest version of the MCODE plug-in for Cytoscape – finds clusters or highly interconnected regions in any network loaded into Cytoscape (14) freely available at <http://cbio.mskcc.org/~bader/software/mcode/>.

4. A modern version of the Perl programming language installed on the Linux or UNIX workstation freely available at <http://www.perl.com/>.
5. The A-GLAM package (8) freely available at <ftp://ftp.ncbi.nih.gov/pub/spouge/papers/archive/AGLAM/>.

3. Methods

The size of co-expression networks depends on the number of nodes in the network and the threshold used to define an edge between two nodes. There are a number of distance measures that are often used to compare gene expression profiles (16).

Here we use the Pearson correlation coefficient (PCC) as a metric to measure the similarity between expression profiles and to construct gene co-expression networks (17, 18). We establish a link by an edge between two genes, represented by nodes, if the PCC value is higher or equal to 0.7; this is an arbitrary cut-off that can be adjusted depending on the dataset used. The microarray dataset used here is the yeast cell-cycle progression experiment from Cho et al. (9) and Spellman et al. (10). The semantic similarity method (11) was used to quantitatively assess the functional relationships between *S. cerevisiae* genes.

The A-GLAM software package uses a Gibbs sampling algorithm to identify functional motifs (such as TFBSs, mRNA splicing control elements, or signals for mRNA 3'-cleavage and polyadenylation) in a set of sequences. Gibbs sampling (or more descriptively, successive substitution sampling) is a respected Markov-chain Monte Carlo procedure for discovering sequence motifs (19). Briefly, A-GLAM takes a set of sequences as input. The Gibbs sampling step in A-GLAM uses simulated annealing to maximize an 'overall score', a figure of merit corresponding to a Bayesian marginal log-odds score. The overall score is given by

$$s = \sum_{i=1}^w \left(\log_2 \frac{(a-1)!}{(c+a-1)!} + \sum_{(j)} \left\{ \log_2 \left[\frac{(c_{ij}+a_j-1)!}{(a_j-1)!} \right] - c_{ij} \log_2 p_j \right\} \right). \quad [1]$$

In Eq. [1], $m! = m(m-1) \dots 1$ denotes a factorial; a_j , the pseudo-counts for nucleic acid j in each position; $a = a_1 + a_2 + a_3 + a_4$, the total pseudo-counts in each position; c_{ij} , the count of nucleic acid j in position i ; and $c = c_{i1} + c_{i2} + c_{i3} + c_{i4}$, the total number of aligned windows, which is independent of the position i . The rationale behind the overall score s in A-GLAM is explained in detail elsewhere (8).

To initialize its annealing maximization, A-GLAM places a single window of arbitrary size and position at every sequence, generating a gapless multiple alignment of the windowed subsequences. It then proceeds through a series of iterations; on each iteration step, A-GLAM proposes a set of adjustments to the alignment. The proposal step is either a repositioning step or a resizing step. In a repositioning step, a single sequence is chosen uniformly at random from the alignment; and the set of adjustments include all possible positions in the sequence where the alignment window would fit without overhanging the ends of the sequence. In a resizing step, either the right or the left end of the alignment window is selected; and the set of proposed adjustments includes expanding or contracting the corresponding end of all alignment windows by one position at a time. Each adjustment leads to a different value of the overall score s . Then, A-GLAM accepts one of the adjustments randomly, with probability proportional to $\exp(s/T)$. A-GLAM may even exclude a sequence if doing so would improve alignment quality. The temperature T is gradually lowered to $T = 0$, with the intent of finding the gapless multiple alignment of the windows maximizing s . The maximization implicitly determines the final window size. The randomness in the algorithm helps it avoid local maxima and find the global maximum of s . Due to the stochastic nature of the procedure, finding the optimum alignment is not guaranteed. Therefore, A-GLAM repeats this procedure ten times from different starting points (ten runs). The idea is that if several of the runs converge to the same best alignment, the user has increased confidence that it is indeed the optimum alignment. The steps (below) corresponding to E -values and post-processing were then carried out with the PSSM corresponding to the best of the ten scores s .

The individual score and its E -value in A-GLAM: The Gibbs sampling step produces an alignment whose overall score s is given by Eq. [1]. Consider a window of length w that is about to be added to A-GLAM's alignment. Let $\delta_i(j)$ equal 1 if the window has nucleic acid j in position i , and 0 otherwise. The addition of the new window changes the overall score by

$$\Delta s = \sum_{i=1}^w \sum_{(j)} \delta_i(j) \left\{ \log_2 \left[\left(\frac{c_{ij} + a_j}{c + a} \right) / p_j \right] \right\}. \quad [2]$$

The score change corresponds to scoring the new window according to a position-specific scoring matrix (PSSM) that assigns the 'individual score'

$$s_i(j) = \log_2 \left[\left(\frac{c_{ij} + a_j}{c + a} \right) / p_j \right] \quad [3]$$

to nucleic acid j in position i . Equation [3] represents a log-odds score for nucleic acid j in position i under an alternative hypothesis with probability $(c_{ij} + a_j)/(c + a)$ and a null hypothesis with

probability p_{ij} . PSI-BLAST (20) uses Eq. [3] to calculate E -values. The derivation through Eq. [2] confirms the PSSM in Eq. [3] as the natural choice for evaluating individual sequences.

The assignment of an E -value to a subsequence with a particular individual score is done as follows: consider the alignment sequence containing the subsequence. Let n be the sequence length, and recall that w is the window size. If ΔS_i denotes the quantity in Eq. [2] if the final letter in the window falls at position i of the alignment sequence, then $\Delta S^* = \max\{\Delta S_i : i = w, \dots, n\}$ is the maximum individual score over all sequence positions i . We assigned an E -value to the actual value $\Delta S^* = \Delta s^*$, as follows. Staden's method (21) yields $\mathbb{P}\{\Delta S_i \Delta s^*\}$ (independent of i) under the null hypothesis of bases chosen independently and randomly from the frequency distribution $\{p_j\}$. The E -value $E = (n - w + 1)\mathbb{P}\{\Delta S_i \Delta s^*\}$ is therefore the expected number of sequence positions with an individual score exceeding Δs^* . The factor $n - w + 1$ in E is essentially a multiple test correction.

More recently, the A-GLAM package has been improved to allow the identification of multiple instances of an element within a target sequence (22). The optional 'scanning step' after Gibbs sampling produces a PSSM given by Eq. [3]. The new scanning step resembles an iterative PSI-BLAST search based on the PSSM. First, it assigns an 'individual score' to each subsequence of appropriate length within the input sequences using the initial PSSM. Second, it computes an E -value from each individual score to assess the agreement between the corresponding subsequence and the PSSM. Third, it permits subsequences with E -values falling below a threshold to contribute to the underlying PSSM, which is then updated using the Bayesian calculus. A-GLAM iterates its scanning step to convergence, at which point no new subsequences contribute to the PSSM. After convergence, A-GLAM reports predicted regulatory elements within each sequence in the order of increasing E -values; users then have a statistical evaluation of the predicted elements in a convenient presentation. Thus, although the Gibbs sampling step in A-GLAM finds at most one regulatory element per input sequence, the scanning step can now rapidly locate further instances of the element in each sequence.

3.1. Co-expression Network Construction

1. The yeast cell-cycle-regulated expression data are obtained from <http://cellcycle-www.stanford.edu/> (see Note 1).
2. Pairwise Pearson correlation coefficient (PCC) values are calculated using a subroutine implemented in the Perl programming language (23) (see Note 2).
3. The co-expression network is constructed with all gene pairs with a PCC greater or equal to 0.7 and is formatted according to the simple interaction file (SIF) described in the Cytoscape manual available at <http://www.cytoscape.org/> (see Note 3).

4. The co-expression network can be loaded in Cytoscape, which is an open-source software for integrating biomolecular interaction networks. Cytoscape is available for a variety of operating systems, including Windows, Linux, Unix, and Mac OS X.

3.2. Functional Similarity Network Construction

1. Gene ontology (GO) annotations for yeast gene products come from the *Saccharomyces* Genome Database (SGD) and were downloaded from http://www.geneontology.org/cgi-bin/downloadGOGA.pl/gene_association.sgd.gz. The evidence supporting such annotations is captured by evidence codes, including TAS (Traceable Author Statement) and IEA (Inferred from Electronic Annotation). While TAS refers to peer-reviewed papers and indicates strong evidence, IEA denotes automated predictions, not curated by experts, i.e., generally less reliable annotations. For this reason, IEA annotations were excluded from this study.
2. Functional relationships between *S. cerevisiae* genes were assessed quantitatively using a semantic similarity method (11) based on the gene ontology (GO). We first computed semantic similarity among GO terms from the *Biological Process* hierarchy using the Lin metric. This metric is based on information content and defines **term-term similarity**, i.e., the semantic similarity $\text{sim}(c_i, c_j)$ between two terms c_i and c_j as

$$\text{sim}(c_i, c_j) = \frac{2 \times \max_{c \in S(c_i, c_j)} [\log(p(c))]}{\log(p(c_i)) + \log(p(c_j))}, \quad [4]$$

where $S(c_i, c_j)$ represents the set of ancestor terms shared by both c_i and c_j , ‘max’ represents the maximum operator, and $p(c)$ is the probability of finding c or any of its descendants in the SGD database. It generates normalized values between 0 and 1. **Gene-gene similarity** results from the aggregation of term-term similarity values between the annotation terms of these two genes. In practice, given a pair of gene products, g_k and g_p , with sets of annotations A_k and A_p comprising m and n terms, respectively, the gene-gene similarity, $\text{SIM}(g_k, g_p)$, is defined as the *highest average* (inter-set) *similarity* between terms from A_i and A_j :

$$\text{SIM}(g_i, g_j) = \frac{1}{m+n} \times \left\{ \sum_k \max_p [\text{sim}(c_k, c_p)] + \sum_p \max_k [\text{sim}(c_k, c_p)] \right\}, \quad [5]$$

where $\text{sim}(c_i, c_j)$ may be calculated using **Eq. [1]**. This aggregation method (12) can be understood as a variant of the Dice similarity.

3. The functional similarity network is constructed using semantic similarity greater or equal to 0.7 and is formatted according to the simple interaction file (SIF).

4. Functional relationships in a group of genes can be further explored in Cytoscape using the BiNGO plug-in (24). Here we have used the hypergeometric test to assess the statistical significance ($p < 0.05$) and the Benjamini & Hochberg False Discovery Rate (FDR) correction (25).

3.3. Intersection Network Construction

1. The yeast co-expression and functional similarity networks are loaded in Cytoscape and the intersection network can be obtained by using the Graph Merge plug-in, freely available at the Cytoscape Web site. The nodes that are connected by having similar expression profiles and GO annotations are present in the intersection network (**Fig. 1.1**) (*see Note 4*).

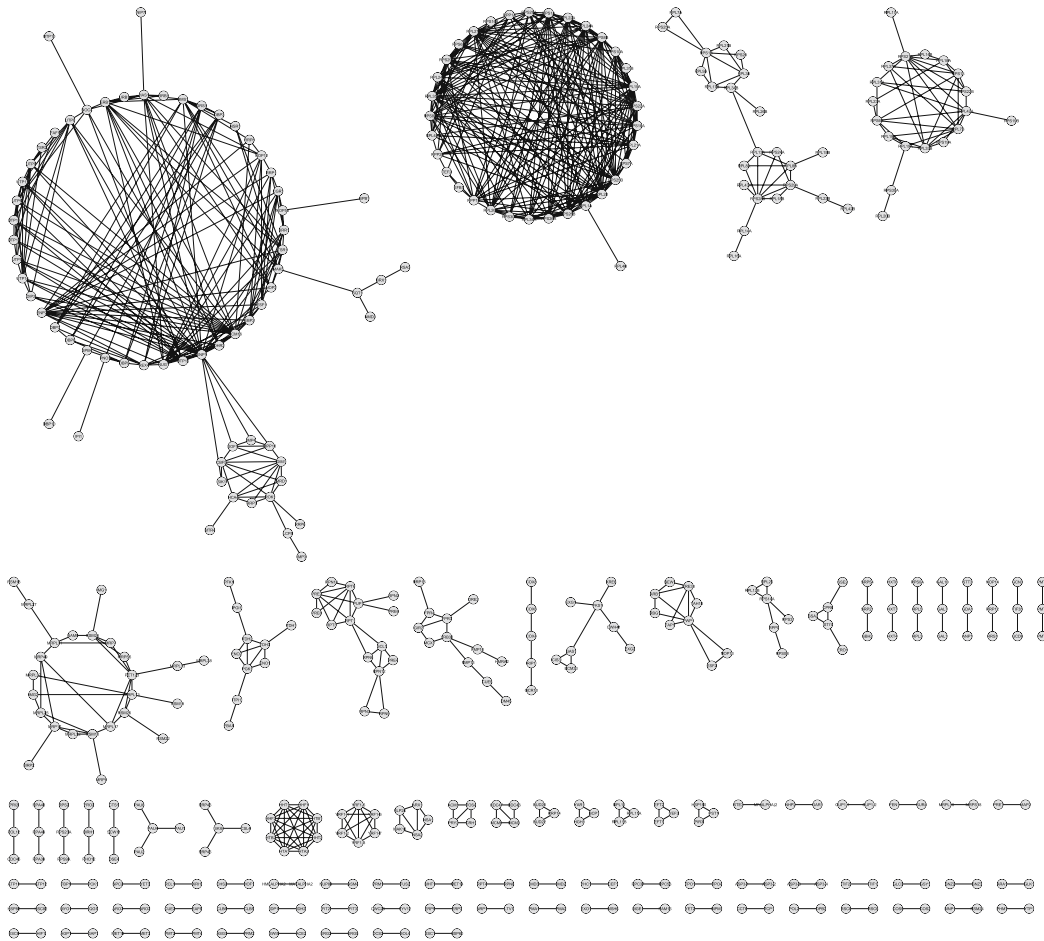


Fig. 1.1. Yeast cell-cycle gene co-expression and GO intersection network. The intersection network topology is shown for yeast genes, represented by nodes linked by one or more edges as described in the text. An edge represents both co-expression and functional linkage between the nodes connected.

2. The intersection network can be visualized using a variety of layouts in Cytoscape. A circular layout of the intersection network using the *yFiles Layouts* plug-in is depicted in **Fig. 1.1**.

3.4. Identification of Highly Interconnected Regions

1. The identification of dense gene clusters in the intersection network is done using the MCODE Cytoscape plug-in (14) (see **Note 5**). The clusters identified share similar expression patterns and functions as described by GO (**Fig. 1.2**).

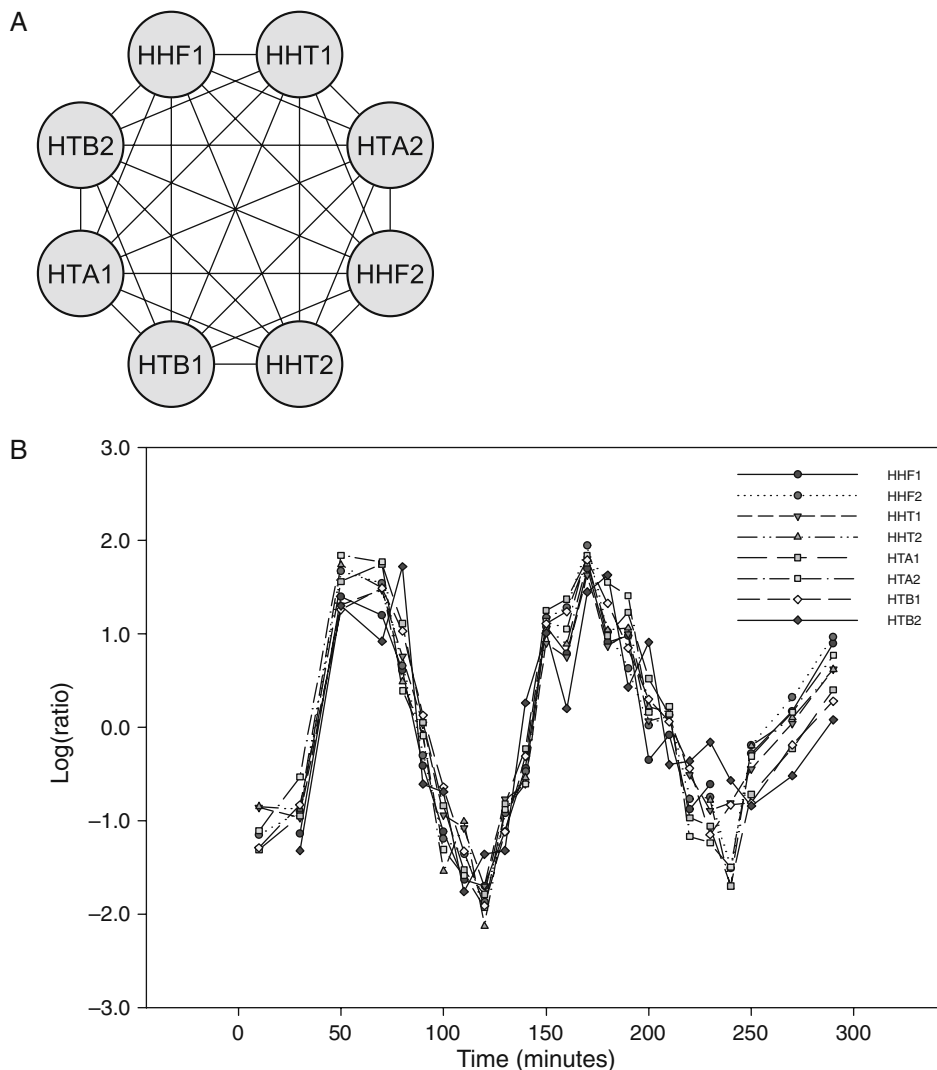


Fig. 1.2. Core histone gene cluster in the intersection network. **A.** Highly connected cluster identified by MCODE corresponds to eight core histone genes present in the yeast genome. The eight nodes are connected by 28 co-expression and functional edges. **B.** Expression profiles of the core histone genes over the cell cycle. **C.** Over-represented GO terms in the Biological Process category for the core histone genes. The statistical significance of each GO term is related to the intensity of the colored circles (see **Note 5**).

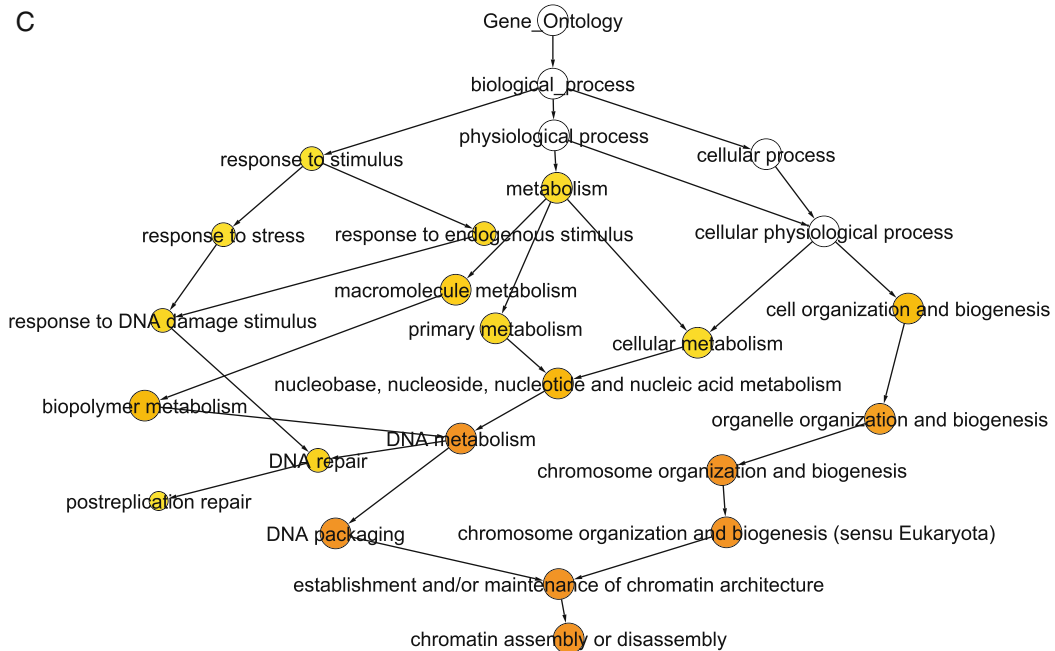


Fig. 1.2. (continued)

3.5. Identification of Proximal Promoter Regions

1. The *Saccharomyces* Genome Database (SGD) maintains the most current annotations of the yeast genome (see <http://www.yeastgenome.org/>). The SGD FTP site contains the DNA sequences annotated as intergenic regions in FASTA format (available at ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/intergenic/), indicating the 5' and 3' flanking features. Additionally, a tab-delimited file with the annotated features of the genome is necessary to determine the orientation of the intergenic regions relative to the genes (available at ftp://genome-ftp.stanford.edu/pub/yeast/chromosomal_feature/). The two files can be used to extract upstream intergenic regions (26) for the genes present in the intersection network clusters (see Note 6).

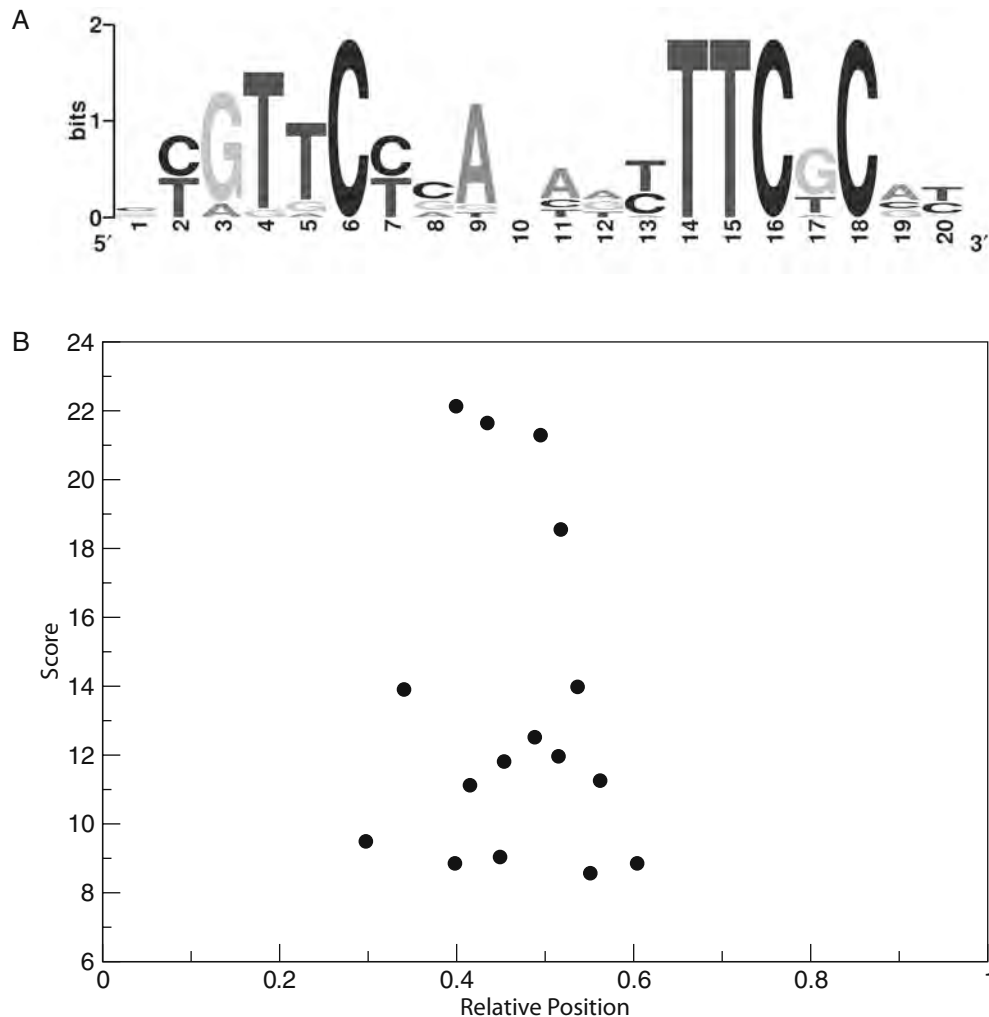
3.6. Identification of cis-Regulatory Elements in Promoter Regions

1. Construct FASTA files for each of the gene clusters identified by MCODE.
2. Install the A-GLAM package (see Note 7).
3. The A-GLAM package has a number of options that can be used to adjust search parameters (see Note 8).
\$ aglam
Usage summary: aglam [options] myseqs.fa

Options:

- h help: print documentation
- n end each run after this many iterations without improvement (10,000)
- r number of alignment runs (10)
- a minimum alignment width (3)
- b maximum alignment width (10,000)
- j examine only one strand
- i word seed query ()
- f input file containing positions of the motifs ()
- z turn off ZOOPS (force every sequence to participate in the alignment)
- v print all alignments in full
- e turn off sorting individual sequences in an alignment on p-value
- q pretend residue abundances = 1/4
- d frequency of width-adjusting moves (1)
- p pseudocount weight (1.5)
- u use uniform pseudocounts: each pseudocount = p/4
- t initial temperature (0.9)
- c cooling factor (1)
- m use modified Lam schedule (default = geometric schedule)
- s seed for random number generator (1)
- w print progress information after each iteration
- l find multiple instances of motifs in each sequence
- k add instances of motifs that satisfy the cutoff e-value (0)
- g number of iterations to be carried out in the post-processing step (1,000)

4. Run A-GLAM to identify the regulatory elements present in the gene clusters with similar expression patterns and GO annotations (*see Note 9*). A-GLAM correctly identifies an experimentally characterized element known to regulate core histone genes in yeast (27). The alignments produced by A-GLAM can be represented by sequence logos (28, 29) and the positional preferences of the elements can be evaluated by plotting the score against relative positions, normalized by sequence length, in the promoter sequences (**Fig. 1.3**).



4. Notes



1. The yeast cell cycle data from the Web site include the experiments from Cho et al. (9) and Spellman et al. (10).
2. The following Perl code can be used to calculate the PCC:

```
my$r = correlation(\@{$values{$probe1}}, \@{$values
{$probe2}});
```

```

sub covariance {
  my ($array1ref,$array2ref) = @_;
  my ($i,$result);
  for ($i = 0;$i < @$array1ref;$i++) {$result += $array1
    ref->[$i] * $array2ref->[$i];
  }
  $result /= @$array1ref;
  $result -= mean($array1ref) * mean($array2ref);
}

sub correlation {
  my ($array1ref,$array2ref) = @_;
  my ($sum1,$sum2);
  my ($sum1_squared, $sum2_squared);
  foreach (@$array1ref) {$sum1 += $_;$sum1_squared
    += $_ ** 2 }
  foreach (@$array2ref) {$sum2 += $_;$sum2_squared
    += $_ ** 2 }
  return (@$array1ref ** 2) * covariance($array1ref,
    $array2ref) / sqrt(((@$array1ref * $sum1_squared) -
    ($sum1 ** 2)) * ((@$array1ref * $sum2_squared) -
    ($sum2 ** 2)));
}

sub mean {
  my ($arrayref) = @_;
  my $result;
  foreach (@$arrayref) {$result += $_ }
  return $result / @$arrayref;
}

```

3. The simple interaction file (SIF or .sif format) consists of lines where each node, representing a protein, is connected by an edge to a different protein in the network. Lines from the simple interaction file from the co-expression network:

```

RPL12A pp THR1
RPL12A pp TIF2
RPL12A pp TIF1
RPL12A pp GUK1
RPL12A pp URA5
RPL12A pp RPL1B
RPL12A pp SSH1
RPL12A pp SNU13
RPL12A pp RPL23B
SHU1 pp DON1

```

Two nodes are connected by a relationship type that in this case is pp. The nodes and their relationships are delimited by a space or a tab (see the Cytoscape manual for more detailed information).

4. Two or more networks can be used to calculate their intersection as needed to select for connections that meet certain criteria. The researcher can overlay protein–protein interactions, co-expression and functional networks to identify the protein complexes created under specific experimental conditions.
5. The MCODE plug-in ranks the clusters according to the average number of connections per protein in the complex (Score). The top five clusters identified by MCODE in the intersection network are shown below:

Cluster	Score	Proteins	Interactions
1	6.6	15	99
2	3.5	8	28
3	2.267	15	34
4	2	5	10
5	2	5	10

The BiNGO plug-in can be used to determine the GO terms statistically over-represented in a group of genes. Here we show the results for cluster 2:

Selected statistical test : Hypergeometric test

Selected correction : Benjamini & Hochberg False Discovery Rate (FDR) correction

Selected significance level : 0.05

Testing option : Test cluster versus complete annotation

The selected cluster :

HHT1 HHF1 HTA1 HHT2 HHF2 HTA2 HTB1 HTB2

Number of genes selected : 8

Total number of genes in annotation : 5932

6. There are a number of Web sites that facilitate the extraction of promoter sequences. A service for the extraction of human, mouse, and rat promoters is freely available at <http://bio.wulf.bu.edu/zlab/promoser/>
7. The A-GLAM package is currently available in source code and binary forms for the Linux operating system (see <ftp://ftp.ncbi.nih.gov/pub/spouge/papers/archive/AGLAM/>).

GO ID	<i>P</i> -value	Corrected <i>P</i> -value	Description
6333	4.9168E-15	1.2292E-13	Chromatin assembly or disassembly
6325	2.2510E-12	1.8758E-11	Establishment and/or maintenance of chromatin architecture
6323	2.2510E-12	1.8758E-11	DNA packaging
7001	2.0415E-10	1.2759E-9	Chromosome organization and biogenesis (sensu Eukaryota)
51276	2.5897E-10	1.2949E-9	Chromosome organization and biogenesis
6259	5.9413E-9	2.4756E-8	DNA metabolism
6996	6.9565E-7	2.4845E-6	Organelle organization and biogenesis

Installation of the Linux binary: Get the executable from the FTP site and set execute permissions.

```
$chmod +x aglam
```

Installation from source: Unpack the glam archive and compile A-GLAM.

```
$tar -zxvf aglam.tar.gz
```

```
$cd aglam
```

```
$make aglam
```

8. Possible scenarios and options to modify A-GLAM's behavior.

```
$aglam <myseqs.fa>
```

This command simply uses the standard Gibbs sampling procedure to find sequence motifs in “myseqs.fa”.

```
$aglam <myseqs.fa> -n 20000 -a 5 -b 15 -j
```

This tells the program to search only the given strand of the sequences to find motifs of length between 5 and 15 bp. The flag *n* specifies the number of iterations performed in each of the ten runs. Low values of *n* are adequate when the problem size is small, i.e., when the sequences are short and more importantly there are few of them, but high values of *n* are needed for large problems. In addition, smaller values of *n* are sufficient when there is a strong alignment to be found, but larger values are necessary when there is not, e.g., for finding the optimal alignment of random sequences. You will have to

choose n on a case-by-case basis. This parameter also controls the tradeoff between speed and accuracy.

```
$aglam <myseqs.fa> -i TATA
```

This important option sets the program to run in a “seed”-oriented mode. The above command restricts the search to sequences that are TATA-like. Unlike the procedure followed in the standard Gibbs sampling algorithm, however, A-GLAM continues to align one exact copy of the “seed” in all “seed sequences”. Therefore, A-GLAM uses the seed sequences to direct its search in the remaining non-seed “target sequences”. Using this option leads to the global optimum quickly.

```
$aglam <myseqs.fa> -f <positions.dat>
```

The above command uses an extra option that allows A-GLAM to take a set of positions from an input file “positions.dat”. Like with the “-” flag, this option provides “seeds” for the A-GLAM alignment. Using this command restricts the Gibbs sampling step to aligning the original list of windows specified by the positions in the file. The seed sequences then direct the search in the remaining non-seed sequences.

```
$aglam <myseqs.fa> -l -k 0.05 -g 2000
```

Usable only with version 1.1. This tells the program to find multiple motif instances in each input sequence, via the scanning step (described above). Those instances that receive an E -value less than 0.05 are included in the PSSM. The search for multiple motifs is carried on until either (a) no new motifs are present or (b) the user-specified number of iterations (in this case, it is 2,000) is attained, whichever comes first.

9. A-GLAM uses sequences in FASTA format as input. Cluster number 2, identified by MCODE, is composed of eight genes linked by 28 co-expression and GO connections. Interestingly, the intergenic regions of the same cluster are shared between the genes in the cluster:

```
>B:235796-236494, Chr 2 from 235796-236494,
between YBL003C and YBL002W
TATATATTAAATTTGCTCTTGTTCTGTACTTTCCTAATTCTTATGTA
AAAAGACAAGAAT
TTATGATACTATTTAATAACAAAAACTACCTAAGAAAAGCATCATGCAG
TCGAAATTGA
AATCGAAAAGTAAAACTTTAACGGAACATGTTTGAAATTCTAAGAAAGC
ATACATCTTCA
TCCCTTATATATAGAGTTATGTTTGATATTAGTAGTCATGTTGTAATCT
CTGGCCTAAGT
ATACGTAACGAAAATGGTAGCACGTCGCGTTTATGGCCCCAGGTTAAT
GTGTTCTCTGA
AATTCGCATCACTTTGAGAAATAATGGGAACACCTTACGCGTGAGCTGT
GCCACCGCTT
CGCCTAATAAAGCGGTGTTCTCAAAATTTCTCCCCGTTTTTCAGGATCAC
GAGCGCCATCT
```

AGTTCTGGTAAATCGCGCTTACAAGAACAAAGAAAAGAAACATCGCGT
 AATGCAACAGT
 GAGACACTTGCCGTCATATATAAGGTTTGGATCAGTAACCGTTATTTG
 AGCATAACACA
 GGTTTTTAAATATATTATTATATATCATGGTATATGTGTAAATTTTTT
 TGCTGACTGGT
 TTTGTTTATTTATTTAGCTTTTTAAAAATTTTACTTTCTTCTGTTAAT
 TTTTTCTGATT
 GCTCTATACTCAAACCAACAACAACCTACTCTACAACATA
 >D:914709-915525, Chr 4 from 914709-915525, between
 YDR224C and YDR225W
 TGTATGTGTGTATGGTTTATTTGTGGTTTGACTTGCTATATAGGATAA
 ATTTAATATAA
 CAATAATCGAAAATGCGGAAAGAGAAACGTCTTTAATAAATCTGACCAT
 CTGAGATGATC
 AAATCATGTTGTTTATATACATCAAGAAAACAGAGATGCCCCTTTCTTA
 CCAATCGTTAC
 AAGATAACCAACCAAGGTAGTATTTGCCACTACTAAGGCCAATTCTCTT
 GATTTTAAATC
 CATCGTTCTCATTTTTTCGCGGAAGAAAGGGTGCAACGCGCGAAAAAGT
 GAGAACAGCCT
 TCCCTTTTCGGGCGACATTGAGCGTCTAACCATAGTTAACGACCCAACCG
 CGTTTTCTTCA
 AATTTGAACTCGCCGAGCTCACAAATAATTCATTAGCGCTGTTCCAAAA
 TTTTCGCCTCA
 CTGTGCGAAGCTATTGGAATGGAGTG
 TATTTGGTGGCTCAAAAAAGAGCACAAATAGTTA
 ACTCGTCGTTGTTGAAGAAACGCCGTAGAGATATGTGGTTTCTCATGC
 TGTATTTGTT
 ATTGCCCACTTTGTTGATTTCAAAATCTTTTCTCACCCCCCTTCCCCGT
 CACGAAGCCAG
 CCAGTGGATCGTAAATACTAGCAATAAGTCTTGACCTAAAAAATATATA
 AATAAGACTCC
 TAATCAGCTTGTAGATTTTCTGGTCTTGTGAACCATCATCTATTTACT
 TCCAATCTGTA
 CTTCTCTTCTTGATACTACATCATCATACGGATTTGGTTATTTCTCAGT
 GAATAAACAAC
 TTCAAAACAACAATTTTCATACATATAAAATATAAA
 >N:576052-576727, Chr 14 from 576052-576727, between
 YNL031C and YNL030W
 TGTGGAGTGTGTTGCTTGGATCCTTTAGTAAAAGGGGAAGAACAGTTGGAA
 GGGCCAAAGT
 GGAAGTCACAAAACAGTGGTCTTATATAAAAGAACAAGAAAAAGATTATT
 TATATACAAC
 TCGGTCACAAGAAGCAACGCGAGAGAGCACAAACACGCTGTTATCACGCA
 AACTATGTTT
 TGACACCGAGCCATAGCCGTGATTGTGCGTCACATTGGGCGATAATGAAC
 GCTAAATGAC
 CAACTCCCATCCGTAGGAGCCCCTTAGGGCGTGCCAATAGTTTCACGCGC
 TTAATGCGAA
 GTGCTCGGAACGGACAACCTGTGGTCGTTTGGCACCGGGAAAGTGGTACTA
 GACCGAGAGT
 TTCGCATTTGTATGGCAGGACGTTCTGGGAGCTTCGCGTCTCAAGCTTTT
 TCGGGCGCGA

```

AATGCAGACCAGACCAGAACAAAACAACTGACAAGAAGGCGTTTAATTTA
ATATGTTGTT
CACTCGCGCCTGGGCTGTTGTTATTTCGGCTAGATACATACGTGTTTGTCG
GTATGTAGTT
ATATCATATATAAGTATATTAGGATGAGGCGGTGAAAGAGATTTTTTTT
TTTTCGCTTAA
TTTATTCTTTTCTCTATCTTTTTTCCTACATCTTGTTCAAAAGAGTAGC
AAAAACAACAA
TCAATACAATAAAATA
>B:255683-256328, Chr 2 from 255683-256328, between
YBR009C and YBR010W
ATTTTACTATATTATATTTGTTGCTTGTGTTTGTGTTGCTTTAGTAC
TATAGAGTACA
ATAATGCGACGGAAACCATCATATAGAAAAATATCTCGGTATTTATAG
GAAAAAGAATT
AGACCTTTTCCACAACCAATTTATCCATCAAATTGGTCTTTACCCAATG
AATGGGGAAGG
GGGGGTGGCAATTTACCACCGTATTCGCGGGCATTGCTAAAGTAAACA
ACTTCGGTTTT
TACCACTAACCATTATGGGGAGAAGCGCTCGGAACAGTTTTACTATGTG
AAGATGCGAAG
TTTTCAGAACGCGGTTTCCAAATTCGGCGGGGAGATACAAAAAGATTT
TTGCTCTCGTT
CTCACATTTTCGCATTGTCCCATACATTATCGTTCTCACAATTTCTCAC
ATTTCCCTTGCT
CTGCACCTTTGCGATCCTGGCCGTAATATCTCTCCTTGACTTTTAGCGT
GGAAGATAACG
AAATGCCCGGGCGATTTTCTTTTTTGGTACCCTCCACGGCTCCTTGTTG
AAATACATATA
TAAAAGACTGTGTATTCTTCGGGATACATCTCTTCTCAACCTTTTAT
ATTCTTTCTTT
CTAGTTAATAAGAAAAACATCTAACATAAATATATAAACGCAAACA

```

A-GLAM has a number of useful command line options that can be adjusted to improve *ab initio* motif finding; in this example we have restricted the search to motifs no larger than 20 bp.

```
$aglam -b 20 -l 02.fa
```

```
A-GLAM:  Anchored  Gapless  Local  Alignment  of
Multiple
```

```
Sequences Compiled on Jun 2 2006
```

```

Run 1... 11724 iterations
Run 2... 10879 iterations
Run 3... 10878 iterations
Run 4... 10336 iterations
Run 5... 10181 iterations
Run 6... 10637 iterations
Run 7... 10116 iterations
Run 8... 11534 iterations
Run 9... 10097 iterations
Run 10... 10239 iterations

```

```

! The sequence file was[ 02.fa]
! Reading the file took[ 0] secs
! Sequences in file[ 4]
! Maximum possible alignment width[ 1292]
! Score[ 243.4] bits
! Motif Width[ 20]
! Runs[ 10]

! Best possible alignment:
>B:235796-236494, Chr 2 from 235796-236494, between
  YBL003C and YBL002W
365 AGGCGAAGCGGTGGGCACAG 346 - (21.29360)
  (2.820982e-08)
394 GGGAGAAATTTTGAGAACAC 375 - (13.97930)
  (5.205043e-04)
309 ATGCGAATTCAGAGAACAC 290 - (11.12770)
  (5.771870e-03)
314 TTGAGAAATAATGGGAACAC 333 + (9.034960)
  (2.714569e-02)
>D:914709-915525, Chr 4 from 914709-915525, between
  YDR224C and YDR225W
423 GTGCGAAGCTATTGGAATGG 442 + (18.55810)
  (2.256236e-06)
278 GCGCGAAAAAGTGAGAACAG 297 + (13.90430)
  (6.495526e-04)
418 AGGCGAAAATTTTGGAACAG 399 - (12.51460)
  (2.007017e-03)
262 CCGCGAAAAAATGAGAACGA 243 - (9.499530)
  (2.299132e-02)
>N:576052-576727, Chr 14 from 576052-576727,
  between YNL031C and YNL030W
294 ATGCGAAGTGCTCGGAACGG 313 + (21.65330)
  (1.526033e-08)
367 ATGCGAAACTCTCGGTCTAG 348 - (11.95760)
  (2.781407e-03)
399 ACGCGAAGCTCCAGAACGT 380 - (11.25120)
  (5.253971e-03)
288 GCGTGAAACTATTGGCACGC 269 - (8.853600)
  (3.961768e-02)
>B:255683-256328, Chr 2 from 255683-256328, between
  YBR009C and YBR010W
258 GGGAGAAGCGCTCGGAACAG 277 + (22.13350)
  (6.281785e-09)
293 ATGCGAAGTTTTCAGAACGC 312 + (11.81510)
  (3.041439e-03)
409 GTGAGAAATTGTGAGAACGA 390 - (8.852760)
  (3.780865e-02)
375 ATGCGAAAATGTGAGAACGA 356 - (8.564750)
  (4.774790e-02)

```

```

! 16 sequences in alignment
! Residue abundances:Pseudocounts
! A = 0.312544:0.468816 C = 0.187456:0.281184
! G = 0.187456:0.281184 T = 0.312544:0.468816
! Total Time to find best alignment [ 15.87] secs

```

Acknowledgments

The authors would like to thank King Jordan for important suggestions and helpful discussions and Alex Brick for his assistance in obtaining intergenic regions during his internship at NCBI. This research was supported by the Intramural Research Program of the NIH, NLM, NCBI.

References

1. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004, 431(7004): 99–104.
2. Bieda M, Xu X, Singer MA, Green R, Farnham PJ. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res* 2006, 16(5):595–605.
3. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 2004, 116(4): 499–509.
4. Guccione E, Martinato F, Finocchiaro G, Luzi L, Tizzoni L, Dall’Olio V, Zardo G, Nervi C, Bernard L, Amati B. Myc-binding-site recognition in the human genome is determined by chromatin context. *Nat Cell Biol* 2006, 8(7):764–770.
5. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005, 23(1): 137–144.
6. Ohler U, Niemann H. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet* 2001, 17(2): 56–60.
7. Marino-Ramirez L, Spouge JL, Kanga GC, Landsman D. Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res* 2004, 32(3): 949–958.
8. Tharakaraman K, Marino-Ramirez L, Sheettlin S, Landsman D, Spouge JL. Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics* 2005, 21(Suppl 1): i440–448.
9. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 1998, 2(1):65–73.
10. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 1998, 9(12):3273–3297.
11. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 2003, 19(10): 1275–1283.
12. Azuaje F, Wang H, Bodenreider O. Ontology-driven similarity approaches to supporting gene functional assessment. In: *Proceedings of the ISMB’2005 SIG Meeting on Bio-Ontologies*. Detroit, MI, 2005:9–10.

13. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003, 13(11):2498–2504.
14. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, 4:2.
15. Tsaparas P, Marino-Ramirez L, Bodenreider O, Koonin EV, Jordan IK. Global similarity and local divergence in human and mouse gene co-expression networks. *BMC Evol Biol* 2006, 6:70.
16. Babu MM. An introduction to microarray data analysis. In: *Computational Genomics: Theory and Application*, Edited by Grant RP. Cambridge, UK: Horizon Bioscience, 2004:225–249.
17. Jordan IK, Marino-Ramirez L, Koonin EV. Evolutionary significance of gene expression divergence. *Gene* 2005, 345(1): 119–126.
18. Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV. Conservation and coevolution in the scale-free human gene coexpression network. *Mol Biol Evol* 2004, 21(11): 2058–2070.
19. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993, 262(5131):208–214.
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25(17): 3389–3402.
21. Staden R. Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* 1989, 5(2):89–96.
22. Tharakaraman K, Marino-Ramirez L, Sheetlin S, Landsman D, Spouge JL. Scanning sequences after Gibbs sampling to find multiple occurrences of functional elements. *BMC Bioinformatics* 2006, 7:408.
23. Orwant J, Hietaniemi J, Macdonald J. *Mastering Algorithms with Perl*. Sebastopol, CA: O'Reilly, 1999.
24. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 2005, 21(16): 3448–3449.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B* 1995, 57(1):289–300.
26. Marino-Ramirez L, Jordan IK, Landsman D. Multiple independent evolutionary solutions to core histone gene regulation. *Genome biology* 2006, 7(12):R122.
27. Eriksson PR, Mendiratta G, McLaughlin NB, Wolfsberg TG, Marino-Ramirez L, Pompa TA, Jainerin M, Landsman D, Shen CH, Clark DJ. Global regulation by the yeast Spt10 protein is mediated through chromatin structure and the histone upstream activating sequence elements. *Mol Cell Biol* 2005, 25(20):9127–9137.
28. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 1990, 18(20): 6097–6100.
29. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res* 2004, 14(6): 1188–1190.

Chapter 13

Promoter Analysis: Gene Regulatory Motif Identification with A-GLAM

Leonardo Mariño-Ramírez, Kannan Tharakaraman, John L. Spouge, and David Landsman

Abstract

Reliable detection of *cis*-regulatory elements in promoter regions is a difficult and unsolved problem in computational biology. The intricacy of transcriptional regulation in higher eukaryotes, primarily in metazoans, could be a major driving force of organismal complexity. Eukaryotic genome annotations have improved greatly due to large-scale characterization of full-length cDNAs, transcriptional start sites (TSSs), and comparative genomics. Regulatory elements are identified in promoter regions using a variety of enumerative or alignment-based methods. Here we present a survey of recent computational methods for eukaryotic promoter analysis and describe the use of an alignment-based method implemented in the A-GLAM program.

Key Words: Promoter regions, transcription factor binding sites, enumerative methods, promoter comparison.

1. Introduction

The establishment and maintenance of temporal and spatial patterns of gene expression are achieved primarily by transcription regulation. Additionally, the precise control of timing and location of gene expression depends on the interaction between transcription factors and *cis*-acting sequence elements in promoter regions. Transcription factors can induce or repress gene expression upon binding of their cognate sequence element in the DNA. The discovery and categorization of the entire collection of transcription factor-binding sites (TFBSs) of an organism are among the greatest challenges in computational biology (1). Large-scale efforts involving genome mapping and identification of TFBS in

lower eukaryotes, such as the yeast *Saccharomyces cerevisiae*, have been successful (2). In contrast, similar efforts in vertebrates have proven difficult due to the presence of repetitive elements and an increased regulatory complexity (3–5).

The accurate prediction and identification of regulatory elements in higher eukaryotes remains a challenge for computational biology, despite recent progress in the development or improvement of different algorithms (6–19). Different strategies for motif recognition have been benchmarked to compare their performance (20). Typically, computational methods for identifying *cis*-regulatory elements in promoter sequences fall into two classes, enumerative and alignment techniques (21). We have developed algorithms that use enumerative approaches to identify *cis*-regulatory elements statistically significant over-represented in promoter regions (22). Subsequently, we developed an algorithm that combines both enumeration and alignment techniques to identify statistically significant *cis*-regulatory elements positionally clustered relative to a specific genomic landmark (23,24).

Promoter identification is the first step in the computational analysis that leads to the prediction of regulatory elements. In lower Eukaryotes this is a rather simple problem due to a relative high gene density with respect to the genome size. The yeast *Saccharomyces cerevisiae* has ~70% of its genome coding for proteins and its intergenic regions are fairly short (~440 bp in length) (25). In contrast, the human genome has a relative low gene density, with ~3% of the genome coding for proteins (26); this poses significant challenges for the identification of both the promoter and its regulatory elements. Despite the complexity of gene expression regulation in higher Eukaryotes (27), we now have a number of experimental and computational resources that can assist in the delineation of mammalian promoter regions. The experimental resources include full-length cDNA collections (28) and transcriptional start sites (TSS) (29). Additionally, complementary computational resources include the database of transcriptional start sites (DBTSS) (30) and promoter identification services (31–33). Many regulatory elements are located in the proximal promoter region (PPR) located a few hundred bases upstream the TSS (22) and the PPR can be generally defined by its low transposable element content (34).

The computational methods for the prediction and identification of transcription factor binding sites can be divided in two broad categories: algorithms for *de novo* identification and algorithms that recognize elements using prior knowledge of the elements. Enumerative and alignment methods form part of the *de novo* algorithms. Enumerative algorithms use exhaustive methods to examine exact DNA words of a fixed length to rank them according to a specific function that determine over-representation relative to a background distribution. An enumerative

method that estimates p -values with the standard normal approximation associated with z -scores (22) has been successfully applied for the identification of regulatory elements in higher Eukaryotes (35). Other enumerative methods include Weeder (16, 17), oligonucleotide frequency analysis (36), and QuickScore (14).

Alignment methods aim to identify functional elements by a multiple local alignment of all sequences of interest. The most popular algorithms in this category use an optimization procedure based in probabilistic sequence models to find statistically significant motifs; these include Gibbs sampling (37) or expectation maximization (11). Approaches that use a combination of enumerative and alignment methods have shown a significant improvement in the identification of regulatory elements in promoter sequences (23, 24).

Algorithms that use prior knowledge of known motifs often use position frequency matrices (PFMs) that contain the number of observed nucleotides at each position (38). Methods that assess statistical over-representation of known motifs in a set of sequences have been particularly successful (9). Additionally, motif scores determined by over-representation can be used as a proxy to perform promoter comparisons (39).

2. Program Usage

2.1. The A-GLAM Algorithm

The A-GLAM software package uses a Gibbs sampling algorithm to identify functional motifs in a set of sequences. Gibbs sampling, also known as successive substitution sampling, is a well-known Markov-chain Monte Carlo procedure for discovering sequence motifs (37). In brief, A-GLAM takes a set of sequences in FASTA format as input. The Gibbs sampling step in A-GLAM uses simulated annealing to maximize an “overall score,” corresponding to a Bayesian marginal log-odds score. The overall score is given by

$$s = \sum_{i=1}^w \left(\log_2 \frac{(a-1)!}{(c+a-1)!} + \sum_{(j)} \left\{ \log_2 \left[\frac{(c_{ij} + a_j - 1)!}{(a_j - 1)!} \right] - c_{ij} \log_2 p_j \right\} \right) \quad (1)$$

In equation (1), $m! = m(m-1) \dots 1$ denotes a factorial; a_j , the pseudo-counts for nucleic acid j in each position; $a = a_1 + a_2 + a_3 + a_4$, the total pseudo-counts in each position; c_{ij} , the count of nucleic acid j in position i ; and $c = c_{i1} + c_{i2} + c_{i3} + c_{i4}$, the total number of aligned windows, which is independent of the position i . The underlying principle behind the overall score s in A-GLAM is explained in detail elsewhere (23).

The annealing maximization is initialized when A-GLAM places a single window of arbitrary size and position at every sequence, generating a gapless multiple alignment of the windowed subsequences. The program then proceeds through a series of iterations; on each iteration step, A-GLAM proposes a set of adjustments to the alignment. The proposal step is either a repositioning step or a resizing step. In a repositioning step, a single sequence is chosen uniformly at random from the alignment; and the set of adjustments include all possible positions in the sequence where the alignment window would fit without overhanging the ends of the sequence. In a resizing step, either the right or the left end of the alignment window is selected; and the set of proposed adjustments includes expanding or contracting the corresponding end of all alignment windows by one position at a time. Each adjustment leads to a different value of the overall score s . Then, A-GLAM accepts one of the adjustments randomly, with probability proportional to $\exp(s/T)$. A-GLAM may even exclude a sequence if doing so would improve alignment quality. The temperature T is gradually lowered to $T = 0$, with the intent of finding the gapless multiple alignments of the windows maximizing s . The maximization implicitly determines the final window size. The randomness in the algorithm helps it avoid local maxima and find the global maximum of s . However, due to the stochastic nature of the procedure, finding the optimum alignment it is not guaranteed.

In the default mode, A-GLAM repeats the annealing maximization procedure ten times from different starting points (ten runs). The rationale behind this is that if several of the runs converge to the same best alignment, the user has increased confidence that it is indeed the optimum alignment.

The individual score and its E-value in A-GLAM: The Gibbs sampling step produces an alignment whose overall score s is given by equation (1). Consider a window of length w that is about to be added to A-GLAM's alignment. Let $\delta_i(j)$ equal 1 if the window has nucleic acid j in position i , and 0 otherwise. The addition of the new window changes the overall score by

$$\Delta s = \sum_{i=1}^w \sum_{(j)} \delta_i(j) \left\{ \log_2 \left[\left(\frac{c_{ij} + a_j}{c + a} \right) / p_j \right] \right\} \quad (2)$$

The score change corresponds to scoring the new window according to a position specific scoring matrix (PSSM) that assigns the "individual score"

$$s_i(j) = \log_2 \left[\left(\frac{c_{ij} + a_j}{c + a} \right) / p_j \right] \quad (3)$$

to nucleic acid j in position i . Equation (3) represents a log-odds score for nucleic acid j in position i under an alternative hypothesis with probability $(c_{ij} + a_j)/(c + a)$ and a null hypothesis with

probability p_{ij} . PSI-BLAST (40) uses equation (3) to calculate E-values. The derivation through equation (2) confirms the PSSM in equation (3) as the natural choice for evaluating individual sequences.

The assignment of an E-value to a subsequence with a particular individual score is done as follows. Consider the alignment sequence containing the subsequence. Let n be the sequence length, and recall that w is the window size. If ΔS_i denotes the quantity in equation (2) if the final letter in the window falls at position i of the alignment sequence, then $\Delta S^* = \max\{\Delta S_i : i = w, \dots, n\}$ is the maximum individual score over all sequence positions i . We assigned an E-value to the actual value $\Delta S^* = \Delta s^*$, as follows. Staden's method (41) yields $\mathbb{P}\{\Delta S_i \geq \Delta s^*\}$ (independent of i) under the null hypothesis of bases chosen independently and randomly from the frequency distribution $\{p_j\}$. The E-value $E = (n - w + 1)\mathbb{P}\{\Delta S_i \geq \Delta s^*\}$ is therefore the expected number of sequence positions with an individual score exceeding Δs^* . The factor $n - w + 1$ in E is essentially a multiple test correction.

More recently, the A-GLAM package has been improved to allow the identification of multiple instances of an element within a target sequence (24). The optional "scanning step" after Gibbs sampling produces a PSSM given by equation (3). The new scanning step resembles an iterative PSI-BLAST search based on the PSSM (Fig. 13.1). First, it assigns an "individual score" to each subsequence of appropriate length within the input sequences using the initial PSSM. Second, it computes an E-value from each individual score to assess the agreement between the corresponding subsequence and the PSSM. Third, it permits subsequences with E-values falling below a threshold to contribute to the underlying PSSM, which is then updated using the Bayesian calculus. A-GLAM iterates its scanning step to convergence, at which point no new subsequences contribute to the PSSM. After convergence, A-GLAM reports predicted regulatory elements within each sequence in order of increasing E-values; users then have a statistical evaluation of the predicted elements in a convenient presentation. Thus, although the Gibbs sampling step in A-GLAM finds at most one regulatory element per input sequence, the scanning step can now rapidly locate further instances of the element in each sequence.

2.2. Hardware

The minimum hardware requirements are a personal computer with at least 512 MB of random access memory (RAM) connected to the Internet as well as access to a Linux or UNIX workstation where A-GLAM will be installed. The connectivity between the personal computer and the workstation is typically established by the Secure Shell (SSH) protocol, a widely used open source of the protocol available at <http://www.openssh.org/>.

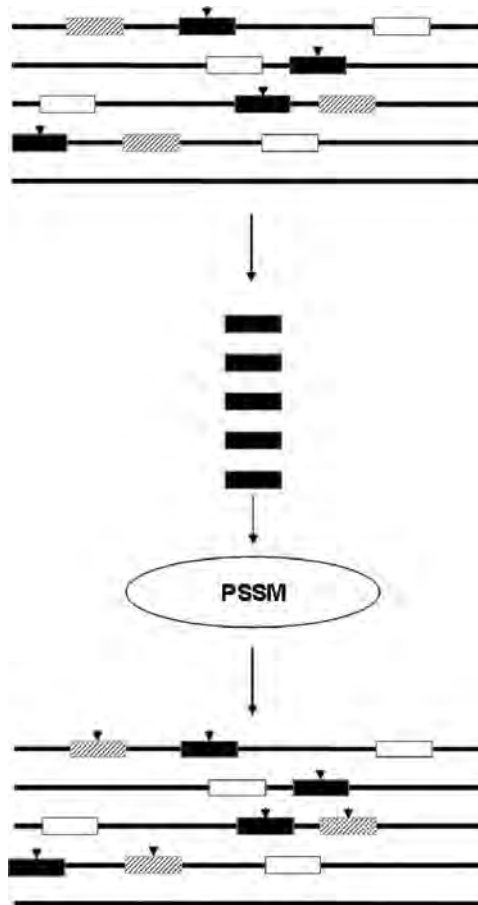


Fig. 13.1. **Strategy to find multiple motif instances in A-GLAM.** The Gibbs sampling identifies up to one motif per sequence (indicated by a *black box* and an *arrowhead*). The sequences are then used to construct a position specific score matrix (PSSM) that is used iteratively to discover multiple motif instances per sequence (indicated by *dashed boxes*).

2.3. Software

A modern version of the Perl programming language installed on the Linux or UNIX workstation freely available at <http://www.perl.com/> will allow the user to parse A-GLAM's output. The A-GLAM package (23) freely available at <http://ftp.ncbi.nih.gov/pub/spouge/papers/archive/AGLAM/> is currently available as source code and binary packages for the Linux operating system.

Installation of the Linux binary: get the executable from the FTP site and set execute permissions.

```
$chmod +x aglam
```

Installation from source: unpack the glam archive in a convenient location and compile A-GLAM.

```
$tar -zxvf aglam.tar.gz
$cd aglam
$make aglam
```

Then you could place the binary in your path: `$HOME/bin` or `/usr/local/bin/`.

2.4. Data Files

A-GLAM accepts input data in FASTA format containing the sequences to be analyzed. The FASTA format consists of one or more sequences identified by a line beginning with the “>” character that should include a unique identifier and a short description about the sequence. The next line(s) should contain the sequence string. A-GLAM expects the standard nucleic acid IUPAC code.

2.5. A-GLAM Options

Some important options to modify A-GLAM’s behavior are described below:

```
$aglam <fasta_file.fa>
```

This command simply uses the standard Gibbs sampling procedure to find sequence motifs in “fasta_file.fa.”

```
$aglam <fasta_file.fa> -n 30000 -a 8 -b 16 -j
```

These sets of commands instruct the program to search only the given strand of the sequences to find motifs of length between 8 and 16 bp. The flag *n* specifies the number of iterations performed in each of the ten runs. Low values of *n* are adequate when the problem size is small, i.e., when the sequences are short and more importantly there are few of them, but high values of *n* are needed for large problems. In addition, smaller values of *n* are sufficient when there is a strong alignment to be found, but larger values are necessary when there is no strong alignment, e.g., for finding the optimal alignment of random sequences. You will have to choose *n* on a case-by-case basis. This parameter also controls the tradeoff between speed and accuracy.

```
$aglam <fasta_file.fa> -i TATA
```

This important option sets the program to run in a “seed” oriented mode. The above command restricts the search to sequences that are TATA-like. Unlike the procedure followed in the standard Gibbs sampling algorithm, however, A-GLAM continues to align one exact copy of the “seed” in all “seed sequences.” Therefore, A-GLAM uses the seed sequences to direct its search in the remaining non-seed “target sequences.” Using this option leads to the global optimum quickly.

```
$aglam <fasta_file.fa> -l -k 0.05 -g 2000
```

Usable only with version 1.1. This set of commands instructs the program to find multiple motif instances in each input sequence via the scanning step (described above). Those instances that receive an E-value less than 0.05 are included in the PSSM. The search for multiple motifs is carried on until either (a) no new motifs are present or (b) the user-specified number of iterations (in this case, it is 2000) is attained, whichever comes first.

3. Example

The A-GLAM package includes documentation and test datasets. Here, we will use a dataset obtained from a large-scale chromatin immunoprecipitation in *Saccharomyces cerevisiae* (2), combined with DNA microarrays (42) to detect interactions between transcription factors and a DNA sequence in vivo. The DNA sequence binding specificity of various transcription factors can then be inferred using A-GLAM on intergenic regions bound by a particular transcription factor. Here, we will use the intergenic regions bound by Snt2p (*see Note 1*).

3.1. Promoter Identification

The *Saccharomyces* Genome Database (SGD) maintains the most current annotations of the yeast genome (*see* <http://www.yeast-genome.org/>). The SGD FTP site contains the DNA sequences annotated as intergenic regions in FASTA format (available at ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/intergenic/), indicating the 5' and 3' flanking features. Additionally, a tab delimited file with the annotated features of the genome is necessary to determine the orientation of the intergenic regions relative to the genes (available at ftp://genome-ftp.stanford.edu/pub/yeast/chromosomal_feature/). The two files can be used to extract upstream intergenic regions. Additionally, there are a number of Web services that facilitate the identification of proximal promoter in mammalian genomes; these include TRED (32), EPD (33), and Promoser (31).

3.2. Identification of cis-Regulatory Elements in Promoter Regions

Construct FASTA files for each of the promoters to be included in the analysis. The Perl programming language can be used in conjunction with BioPerl libraries (freely available at <http://www.bioperl.org/>) to generate files in FASTA format. In this particular example all relevant files can be found on the Fraenkel Web site at http://fraenkel.mit.edu//Harbison/release_v24.

The A-GLAM package has a number of options that can be used to adjust search parameters.

```
$aglam
Usage summary: aglam [options] myseqs.fa
Options:
-h help: print documentation
-n end each run after this many iterations
  without improvement (10000)
-r number of alignment runs (10)
-a minimum alignment width (3)
-b maximum alignment width (10000)
-j examine only one strand
-i word seed query ()
-f input file containing positions of the
  motifs ()
-z turn off ZOOPS (force every sequence to
  participate in the alignment)
-v print all alignments in full
-e turn off sorting individual sequences in an
  alignment on p-value
-q pretend residue abundances = 1/4
-d frequency of width-adjusting moves (1)
-p pseudocount weight (1.5)
-u use uniform pseudocounts: each pseudocount =
  p/4
-t initial temperature (0.9)
-c cooling factor (1)
-m use modified Lam schedule (default = geo-
  metric schedule)
-s seed for random number generator (1)
-w print progress information after each
  iteration
-l find multiple instances of motifs in each
  sequence
-k add instances of motifs that satisfy the
  cutoff e-value (0)
-g number of iterations to be carried out in
  the post-processing step (1000)
```

Run A-GLAM to identify regulatory elements present in the promoter regions bound by Snt2p. A-GLAM uses sequences in FASTA format as input. There are 46 intergenic regions bound by Snt2p that were identified by ChIP-chip in a large-scale study (2). These regions vary in length from 71 to 1,512 bp with an average of 398 bp. A-GLAM is able to identify statistically significant motifs for Snt2p and rank them according to their

individual p -values. A-GLAM has a number of useful command line options that can be adjusted to improve ab initio motif finding; in this example we have restricted the search to motifs no larger than 20 bp and instructed the program to find multiple instances of motifs in each sequence using a strategy that resembles an iterative PSI-BLAST search based on the PSSM constructed by the Gibbs sampling step (24). The output of the A-GLAM program is presented in **Fig. 13.2**. In the default mode, A-GLAM repeats the annealing maximization procedure ten times from different starting points (ten runs). The rationale behind this is that if several of the runs converge to the same best alignment, the user has increased confidence that it is indeed the optimum alignment. The user can adjust the number of alignment runs by setting the `-r` flag (*see Note 2*). The number of iterations can also be adjusted for large datasets. The default value is set at 10,000 without alignment improvement, using the `-n` flag the number of iterations can be increased to extend coverage of the sequence space.

A-GLAM identifies candidate sequences that could serve as Snt2p binding sites. The candidate sequences found by A-GLAM are in agreement with previous findings where other motif finding algorithms were used (2) and **Fig. 13.3**. Additional examples where we have successfully used A-GLAM to complement experimental efforts for the identification of regulatory elements include motifs for Spt10p in yeast and the CREB-binding protein (34, 35). In this particular example, the program constructs a PSSM using the sequences from the optimal alignment to find multiple instances (*see Note 3*). The multiple alignments produced by A-GLAM can be represented graphically by sequence logos (43, 44) (*see Note 4*).

4. Notes



1. The primary data can be obtained from the Fraenkel Laboratory Web site at <http://fraenkel.mit.edu/Harbison/>.
2. The number of alignment runs is 10 by default; however, the user can increase the number of runs to boost the confidence of the results. The user has the option `-v` to print all alignments generated in each run; by default A-GLAM will report only the highest scoring alignment.
3. Alternatively, the user could run A-GLAM without the `-l` flag and construct a position frequency matrix that in turn could be used to scan the target sequences for additional instances of the motif. The TFBS Perl modules for

```

$ aglam -b 20 -l SNT2_YPD.fsa

A-GLAM: Anchored Gapless Local Alignment of Multiple Sequences
Compiled on Feb 9 2008
aglam -l SNT2_YPD.fsa

Run 1... 25340 iterations
Run 2... 26770 iterations
Run 3... 22597 iterations
Run 4... 17786 iterations
Run 5... 23816 iterations
Run 6... 42556 iterations
Run 7... 19556 iterations
Run 8... 22526 iterations
Run 9... 23310 iterations
Run 10... 21531 iterations

! The sequence file was      [SNT2_YPD.fsa]
! Reading the file took      [0] secs
! Sequences in file          [46]
! Maximum possible alignment width [142]
! Score                       [400] bits
! Motif Width                 [12]
! Runs                        [10]

! Best possible alignment:

>iYNL182C 6.2046e-10          202 ATGGCGCTATCA 213 + (10.24060) (1.714353e-02)
>iYBL075C 7.9181e-10          278 GCGGCGCTATCA 267 + (12.62630) (3.136227e-04)
>iYIL160C 1.0190e-09          211 ACGGCGCTATCA 222 + (14.16730) (2.230312e-05)
                                208 AAGGCGCTATCA 197 - (10.97000) (4.259169e-03)
>iYPR183W 1.6110e-09          237 GCGGCGCTATCA 248 + (14.39810) (8.284745e-06)
>iYCR090C-1 2.2463e-09        575 ACGGCGCTATCA 564 - (12.39550) (1.190739e-03)
>iYAL039C-0 5.6844e-09        281 GCGGCGCTATCA 292 + (14.39810) (2.092311e-05)
>iYPR157W 1.0834e-08          343 GTGGCGCTATCA 332 - (10.47150) (1.119393e-02)
                                absent
>iYOL117W 1.2728e-08          252 ATGGCGCTATCA 263 + (12.01250) (1.704126e-03)
>iYLR149C 1.4205e-08          279 GCGGCGCTATCA 268 - (12.62630) (6.283269e-04)
>iYJL093C 3.7648e-08          202 ACGGCGCTATCA 213 + (12.39550) (1.581019e-03)
>iYBR143C 2.6501e-07          221 GTGGCGCTATCA 232 + (12.24330) (1.517043e-03)
>iYLR176C 1.6035e-06          420 ATGGCGCTATCA 431 + (10.24060) (1.333119e-02)
>iYPR104C 6.0302e-06          203 GCGGCGCTAGCA 214 + (12.42200) (4.809407e-04)
>iYBR138C 9.2799e-06          206 CCGCCTCGGCCA 195 - (8.225040) (4.975689e-02)
                                26 CCAAGCTCGCCCC 15 - (8.644490) (1.269803e-02)
>tP(UGG)M 1.4586e-05          99 ACCACTAGACCA 110 + (7.042530) (4.773258e-02)
                                absent
>iYHR217C 1.7438e-05          145 TCGGCGCTATCA 134 - (11.23880) (3.713288e-03)
>iYHR138C 3.9997e-05          absent
>iYKL172W 4.5759e-05          absent
>IntYGL103W 4.7991e-05          absent
>tL(UAG)L2 4.9893e-05          21 ACCACTCGGCCA 10 - (9.819350) (3.647425e-03)
>iYJR152W 5.1753e-05          absent
>tS(AGA)M 6.7229e-05          35 CCTGCGCGGGCA 46 + (9.031130) (6.321172e-03)
>tW(CCA)P 6.8308e-05          81 AAAGCTCTACCA 92 + (10.76890) (1.315551e-03)
>snR128 9.5071e-05          absent
>tI(UAU)L 9.5693e-05          26 GCAACGCGACCG 15 - (8.373710) (1.822708e-02)
>iYCR090C-0 1.0515e-04          absent
>IntYPL081W 1.1828e-04          absent
>SNR190 1.1910e-04          162 CCGATTTCGACCA 151 - (7.925580) (4.569485e-02)
>tL(CAA)C 1.2473e-04          24 ACCGCTCGGCCA 13 - (11.62350) (5.732354e-04)
>tP(AGG)C 1.3420e-04          24 CCGGCTCGCCCC 13 - (9.501020) (4.510700e-03)
>tS(GCU)L 1.9681e-04          absent
>tH(GUG)M 2.0224e-04          absent
>SNR43 2.6811e-04          absent
>tS(CGA)C 3.0856e-04          71 CCAGCGCGGGCA 60 - (10.69280) (1.375755e-03)
>tK(UUU)P 3.1249e-04          55 AACGCTCTACCA 66 + (10.63040) (1.330965e-03)
>tN(GUU)P 4.7144e-04          32 CCAACTTGGCCA 21 - (7.907740) (2.015338e-02)
>tV(AAC)M3 4.8949e-04          60 CCGACTAGACCA 71 + (7.828140) (1.674674e-02)
>tA(UGC)O 5.7662e-04          48 AGCGCGCTATCA 59 + (9.775690) (2.504172e-03)
>tT(AGU)O2 6.8638e-04          60 CCAACTTGGCCA 71 + (7.907740) (1.737360e-02)
>tR(UCU)M2 7.2321e-04          55 GACGCGTTGCCA 66 + (9.845220) (2.902318e-03)
>iYLR228C-1 8.1088e-04          absent
>tQ(UUG)L 8.1134e-04          absent
>tC(GCA)P2 9.1920e-04          46 GCTGCGCTATCA 57 + (11.88000) (2.284798e-04)
>iYDR261C-1 9.4038e-04          absent
>SNR44 9.4060e-04          absent
>tG(GCC)P2 9.5116e-04          26 CCAACGTTGCCA 37 + (9.164880) (5.191352e-03)

! 34 sequences in alignment
! Residue abundances:Pseudocounts
! A = 0.311204:0.466806 C = 0.188796:0.283194 G = 0.188796:0.283194 T = 0.311204:0.466806
! Total Time to find best alignment [13.92] secs

```

Fig. 13.2. **A-GLAM output for a set of sequences containing an SNT2p motif identified using ChIP-chip.** A-GLAM works by analyzing completely random alignment of the sequences and making small refinements over ten alignment runs with many iterations.

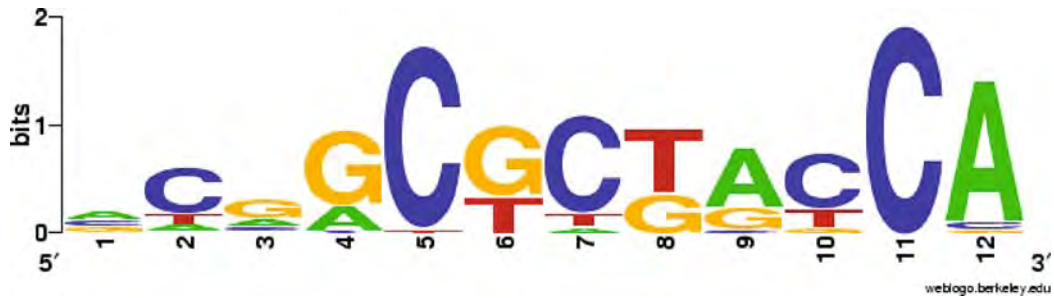


Fig. 13.3. **Snt2p regulatory motif identified with A-GLAM.** Sequence logo representation of the motif obtained from the ungapped multiple sequence alignment identified by A-GLAM.

transcription factor binding detection and analysis provide a flexible and powerful framework (available at <http://tfbs.genereg.net/>).

4. Other Web servers for logo generation include enoLOGOS (available on the Web at <http://biodev.hgen.pitt.edu/enologos/>) and Pictogram (<http://genes.mit.edu/pictogram.html>).

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, NLM, NCBI.

References

1. Elnitski, L., Jin, V. X., Farnham, P. J., and Jones, S. J. (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res* **16**, 1455–64.
2. Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104.
3. Bieda, M., Xu, X., Singer, M. A., Green, R., and Farnham, P. J. (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res* **16**, 595–605.
4. Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509.
5. Guccione, E., Martinato, F., Finocchiaro, G., Luzi, L., Tizzoni, L., Dall’Olio, V., Zardo, G., Nervi, C., Bernard, L., and Amati, B. (2006) Myc-binding-site recognition in the human genome is determined by chromatin context. *Nat Cell Biol* **8**, 764–70.
6. Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* **296**, 1205–14.

7. Workman, C. T., and Stormo, G. D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput* **5**, 467–78.
8. Hertz, G. Z., and Stormo, G. D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–77.
9. Frith, M. C., Fu, Y., Yu, L., Chen, J. F., Hansen, U., and Weng, Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* **32**, 1372–81.
10. Ao, W., Gaudet, J., Kent, W. J., Muttumu, S., and Mango, S. E. (2004) Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* **305**, 1743–6.
11. Bailey, T. L., and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28–36.
12. Eskin, E., and Pevzner, P. A. (2002) Finding composite regulatory patterns in DNA sequences. *Bioinformatics* **18 Suppl 1**, S354–63.
13. Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17**, 1113–22.
14. Régnier, M., and Denise, A. (2004) Rare events and conditional events on random strings. *Discrete Math Theor Comput Sci* **6**, 191–214.
15. Favorov, A. V., Gelfand, M. S., Gerasimova, A. V., Ravcheev, D. A., Mironov, A. A., and Makeev, V. J. (2005) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics* **21**, 2240–5.
16. Pavesi, G., Mereghetti, P., Zambelli, F., Stefani, M., Mauri, G., and Pesole, G. (2006) MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Res* **34**, W566–70.
17. Pavesi, G., Zambelli, F., and Pesole, G. (2007) WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences. *BMC Bioinformatics* **8**, 46.
18. Sinha, S., and Tompa, M. (2003) YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* **31**, 3586–8.
19. Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganier, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D., et al. (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* **16**, 656–68.
20. Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**, 137–44.
21. Ohler, U., and Niemann, H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet* **17**, 56–60.
22. Marino-Ramirez, L., Spouge, J. L., Kanga, G. C., and Landsman, D. (2004) Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res* **32**, 949–58.
23. Tharakaraman, K., Marino-Ramirez, L., Sheetlin, S., Landsman, D., and Spouge, J. L. (2005) Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics* **21 Suppl 1**, i440–8.
24. Tharakaraman, K., Marino-Ramirez, L., Sheetlin, S., Landsman, D., and Spouge, J. L. (2006) Scanning sequences after Gibbs sampling to find multiple occurrences of functional elements. *BMC Bioinformatics* **7**, 408.
25. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., et al. (1996) Life with 6000 genes. *Science* **274**, 546, 563–47.
26. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitz-Hugh, W., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
27. Levine, M., and Tjian, R. (2003) Transcription regulation and animal diversity. *Nature* **424**, 147–51.
28. Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki, D., et al. (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res* **13**, 1273–89.
29. Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R.,

- Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H., et al. (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res* **16**, 55–65.
30. Suzuki, Y., Yamashita, R., Sugano, S., and Nakai, K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res* **32**, D78–81.
 31. Halees, A. S., and Weng, Z. (2004) Promoter: improvements to the algorithm, visualization and accessibility. *Nucleic Acids Res* **32**, W191–4.
 32. Jiang, C., Xuan, Z., Zhao, F., and Zhang, M. Q. (2007) TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res* **35**, D137–40.
 33. Schmid, C. D., Perier, R., Praz, V., and Bucher, P. (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res* **34**, D82–5.
 34. Eriksson, P. R., Mendiratta, G., McLaughlin, N. B., Wolfsberg, T. G., Marino-Ramírez, L., Pompa, T. A., Jainerin, M., Landsman, D., Shen, C. H., and Clark, D. J. (2005) Global regulation by the yeast Spt10 protein is mediated through chromatin structure and the histone upstream activating sequence elements. *Mol Cell Biol* **25**, 9127–37.
 35. Riz, I., Akimov, S. S., Eaker, S. S., Baxter, K. K., Lee, H. J., Marino-Ramírez, L., Landsman, D., Hawley, T. S., and Hawley, R. G. (2007) TLX1/HOX11-induced hematopoietic differentiation blockade. *Oncogene* **26**, 4115–23.
 36. van Helden, J., andre, B., and Collado-Vides, J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* **281**, 827–42.
 37. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–14.
 38. Wasserman, W. W., and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**, 276–87.
 39. Marino-Ramírez, L., Jordan, I. K., and Landsman, D. (2006) Multiple independent evolutionary solutions to core histone gene regulation. *Genome Biol* **7**, R122.
 40. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–402.
 41. Staden, R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* **5**, 89–96.
 42. Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. (2000) Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–9.
 43. Schneider, T. D., and Stephens, R. M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**, 6097–100.
 44. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–90.

Expression Patterns of Protein Kinases Correlate with Gene Architecture and Evolutionary Rates

Aleksey Y. Ogurtsov^{1*}, Leonardo Mariño-Ramírez¹, Gibbes R. Johnson², David Landsman¹, Svetlana A. Shabalina^{1*}, Nikolay A. Spiridonov^{2*}

1 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, **2** Division of Therapeutic Proteins, Center for Drug Evaluation and Research, U. S. Food and Drug Administration, Bethesda, Maryland, United States of America

Abstract

Background: Protein kinase (PK) genes comprise the third largest superfamily that occupy ~2% of the human genome. They encode regulatory enzymes that control a vast variety of cellular processes through phosphorylation of their protein substrates. Expression of PK genes is subject to complex transcriptional regulation which is not fully understood.

Principal Findings: Our comparative analysis demonstrates that genomic organization of regulatory PK genes differs from organization of other protein coding genes. PK genes occupy larger genomic loci, have longer introns, spacer regions, and encode larger proteins. The primary transcript length of PK genes, similar to other protein coding genes, inversely correlates with gene expression level and expression breadth, which is likely due to the necessity to reduce metabolic costs of transcription for abundant messages. On average, PK genes evolve slower than other protein coding genes. Breadth of PK expression negatively correlates with rate of non-synonymous substitutions in protein coding regions. This rate is lower for high expression and ubiquitous PKs, relative to low expression PKs, and correlates with divergence in untranslated regions. Conversely, rate of silent mutations is uniform in different PK groups, indicating that differing rates of non-synonymous substitutions reflect variations in selective pressure. Brain and testis employ a considerable number of tissue-specific PKs, indicating high complexity of phosphorylation-dependent regulatory network in these organs. There are considerable differences in genomic organization between PKs up-regulated in the testis and brain. PK genes up-regulated in the highly proliferative testicular tissue are fast evolving and small, with short introns and transcribed regions. In contrast, genes up-regulated in the minimally proliferative nervous tissue carry long introns, extended transcribed regions, and evolve slowly.

Conclusions/Significance: PK genomic architecture, the size of gene functional domains and evolutionary rates correlate with the pattern of gene expression. Structure and evolutionary divergence of tissue-specific PK genes is related to the proliferative activity of the tissue where these genes are predominantly expressed. Our data provide evidence that physiological requirements for transcription intensity, ubiquitous expression, and tissue-specific regulation shape gene structure and affect rates of evolution.

Citation: Ogurtsov AY, Mariño-Ramírez L, Johnson GR, Landsman D, Shabalina SA, et al. (2008) Expression Patterns of Protein Kinases Correlate with Gene Architecture and Evolutionary Rates. PLoS ONE 3(10): e3599. doi:10.1371/journal.pone.0003599

Editor: Sridhar Hannenhalli, University of Pennsylvania School of Medicine, United States of America

Received: August 7, 2008; **Accepted:** October 9, 2008; **Published:** October 31, 2008

This is an open-access article distributed under the terms of the Creative Commons Public Domain declaration which stipulates that, once placed in the public domain, this work may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose.

Funding: This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nikolay.spiridonov@fda.hhs.gov (NAS); shabalina@ncbi.nlm.nih.gov (SAS)

† These authors contributed equally to this work.

Introduction

Phosphorylation of serine, threonine and tyrosine residues in substrate proteins by protein kinases (PKs) provides a fundamental mechanism for the control of cell division, growth and apoptosis, metabolic activity, adhesion and migration, and mediates cell responses upon environmental stimuli [1,2,3]. At the molecular level, phosphorylation-dephosphorylation allows fast and sensitive regulation of enzyme activity. It is also a major mechanism of transmembrane signal transduction and amplification in the branching network of intracellular PK cascades that ultimately control gene expression by phosphorylation of transcription factors. Phosphorylation of protein substrates creates binding sites for protein domains which recognize specific

phosphorylated amino acid sequences, thereby mediating protein-protein interactions [3,4].

The eukaryotic PK superfamily is subdivided into two broad groups of conventional and atypical kinases. Conventional PKs have been classified into eight families based on the structure and sequence similarities of their conserved eukaryotic catalytic domains. A smaller group of atypical PKs consists of several families that do not carry well conserved kinase domains. Still, many atypical protein kinases show evidence for enzyme activity. The number of PK genes in the animal genome progressively grows from lower to higher organisms, paralleling the evolutionary increase in the total number of genes and the complexity of organization. The protein kinase complement of the human genome (kinome) includes 518 predicted genes, comprising the third-largest gene superfamily [5]. Compar-

ative analysis of the mouse genome performed by different research groups identified 540 to 561 candidate protein kinase genes [6,7]. According to a more recent conservative estimate, the human and the mouse genomes contain 504 and 508 PK genes, correspondingly [8]. The majority of the human protein kinases have orthologs in the mouse, implying similar biological functions in both organisms. Some of these enzymes are restricted to or predominantly expressed in specialized tissues or cell types.

Expression of PK genes is subject to complex transcriptional control, which is not fully understood. Although orthologization and evolutionary conservation of PK protein sequences has been well established, little is known about evolutionary conservation and the function of non-coding DNA sequences of PK genes. Insights into the function of non-coding DNA can be gained from comparative analysis. According to estimations by different authors, fraction of selectively constrained non-coding DNA sequences in mammalian genomes represent from 3% (when highly conserved sequences alone are taken into account) to 10% or more (when weaker conservation is also considered) [9,10]. Evolutionary conservation of non-coding DNA is controlled, at least in part, by negative selection and high interspecies homology of non-coding DNA sequences suggest their important regulatory function. For example, the vast majority of experimentally defined binding sites for the human skeletal muscle-specific transcription factors are confined to the most constrained orthologous sequences in the rodent genome [11]. Because patterns of gene regulation and the corresponding regulatory controls are often conserved between species, cross-species sequence comparison, so-called “phylogenetic footprinting”, may identify functional gene regulatory elements. Alignment algorithms based on interspecies sequence comparison have successfully been used to identify regulatory sites of genes expressed in the skeletal muscle [11] and endothelial tissue [12]. Here we employed a similar approach for identification of regulatory elements in non-coding regions of mammalian PK genes.

We analyzed 497 orthologous genomic loci of the human and mouse PK genes with the total length over 64 Mb, constituting about 2% of a mammalian genome. The goals of the present study were: *i*) evaluation of sequence conservation and evolutionary rates in non-transcribed, transcribed and translated regions of PK genes, *ii*) evaluation of the gene architecture and features of structural domains in differentially expressed genes, *iii*) identification of sequence elements and regulatory signals associated with transcription levels and message abundance, *iv*) evaluation of PK tissue expression patterns and sequence elements associated with tissue-specific gene expression. Here we present data on the relative expression levels and tissue-specific expression for the human kinome. We show that architecture of regulatory PK genes significantly differs from other protein coding genes and explore relationships between gene structure, evolutionary conservation, transcript levels and breadth of gene expression. We demonstrate that architecture of PK genomic loci correlates with the mode of gene expression and proliferative activity of the tissue where these genes are predominantly expressed. We describe evolutionarily conserved signals associated with transcript abundance and tissue-specific expression. Our results suggest that requirements for ubiquitous expression and tissue-specific regulation affect gene structure and impose selection pressure on the protein-coding and non-coding gene regions.

Results

Transcription levels and tissue-specific expression of PK genes

We evaluated expression of PK and non-PK genes based on the numbers of gene-specific ESTs in GenBank originating from

normal human tissues, which reflect mRNA abundance and relative gene transcription levels. The vast majority of PK genes scored moderate EST numbers (84 ESTs average) in contrast with highly transcribed housekeeping non-PK genes (2000 or more ESTs). Based on this analysis, PKs fall into a category of moderately transcribed genes, which is consistent with their regulatory role. These data are in agreement with overall PK expression levels presented in the Gene Expression Atlas. For evaluation of PK expression patterns and the relative abundance of PK messages in different organs, we sorted gene-specific ESTs in accordance with their organ and tissue origins. ESTs originating from the brain and nervous tissue were most numerous in our dataset, followed by ESTs from testis and placenta (Figure 1). Distribution of PK-specific ESTs in 20 normal human organs and tissues is presented in Table S1. Our results showed that the majority of PK genes were broadly expressed in many tissues. At the same time, a number of PK genes showed distinct organ-specific and tissue-specific expression patterns.

Remarkably, EST libraries from nervous and testicular tissues were enriched with PK tags, relative to libraries from other tissues (Figure 1A), suggesting increased phosphorylation-dependent regulation in the brain and testis. Therefore, we focused our attention on PKs up-regulated in these two tissues. A diverse group of protein kinases that includes many members of CAMK1, CAMK2, DCAMKL, Eph, CDK, PKC and other families was predominantly expressed in the brain and nervous tissue (Table S1). Expression of VACAMKL, CaMK2 alpha, EphA7, PKC gamma and PAK5 was effectively restricted to the brain, and many of the nervous tissue-specific PKs scored high numbers of ESTs in GenBank, indicating active transcription. A smaller PK group was preferentially expressed in the testis (BRDT, HIPK4, MISR2, SgK307, SgK396, SSTK, TSSK1, TSSK2, TSSK4 and others). Several genes were predominantly expressed in placenta (TXK, FLT1, ACTR2B), muscle and heart (skMLCK, MSSK1), and other tissues. In the cases where experimental results are available, our identification of tissue-specific protein kinases was supported by data from literature. For example, five testicular protein kinases, TSSK1, TSSK2, SSTK, CAMK4, and Haspin, are specifically expressed in haploid germ cells and two of these enzymes (CAMK4 and SSTK) are indispensable for normal progression of spermatogenesis and male fertility [13,14,15,16]. Experimental evidence for brain- and neuron-specific expression was obtained for VACAMKL [17], PAK5 [18], Eph receptor tyrosine kinase EphA4 [19], CaMK1 gamma [20], CaMK2 alpha [21], PKC gamma [22], CDK5 [23] and some other kinases identified in our search.

Structural features of PK genes associated with expression levels and breadth

For evaluation of gene architecture, we analyzed length and GC content in different gene functional domains in 510 human PK genomic loci. To compare structural properties of PK genes with overall trends for other genes, we used control group of 7,711 well-annotated human non-PK genes. Genomic architecture of PK and non-PK genes significantly differed. As seen from Figure 2, PK genes occupy larger genomic loci, possess significantly longer exons and spacer regions, and encode larger proteins, relative to the group of non-PK genes. PK genes also tend to have more GC-rich UTRs relative to non-PK genes (Table 1). Remarkably, lengths of gene loci, 5'-spacers, introns and UTRs of the human PK genes were ~15% longer than for the mouse PK genes, revealing higher gene complexity (Figure S1). Same trend was observed for non-PK genes (data not shown).

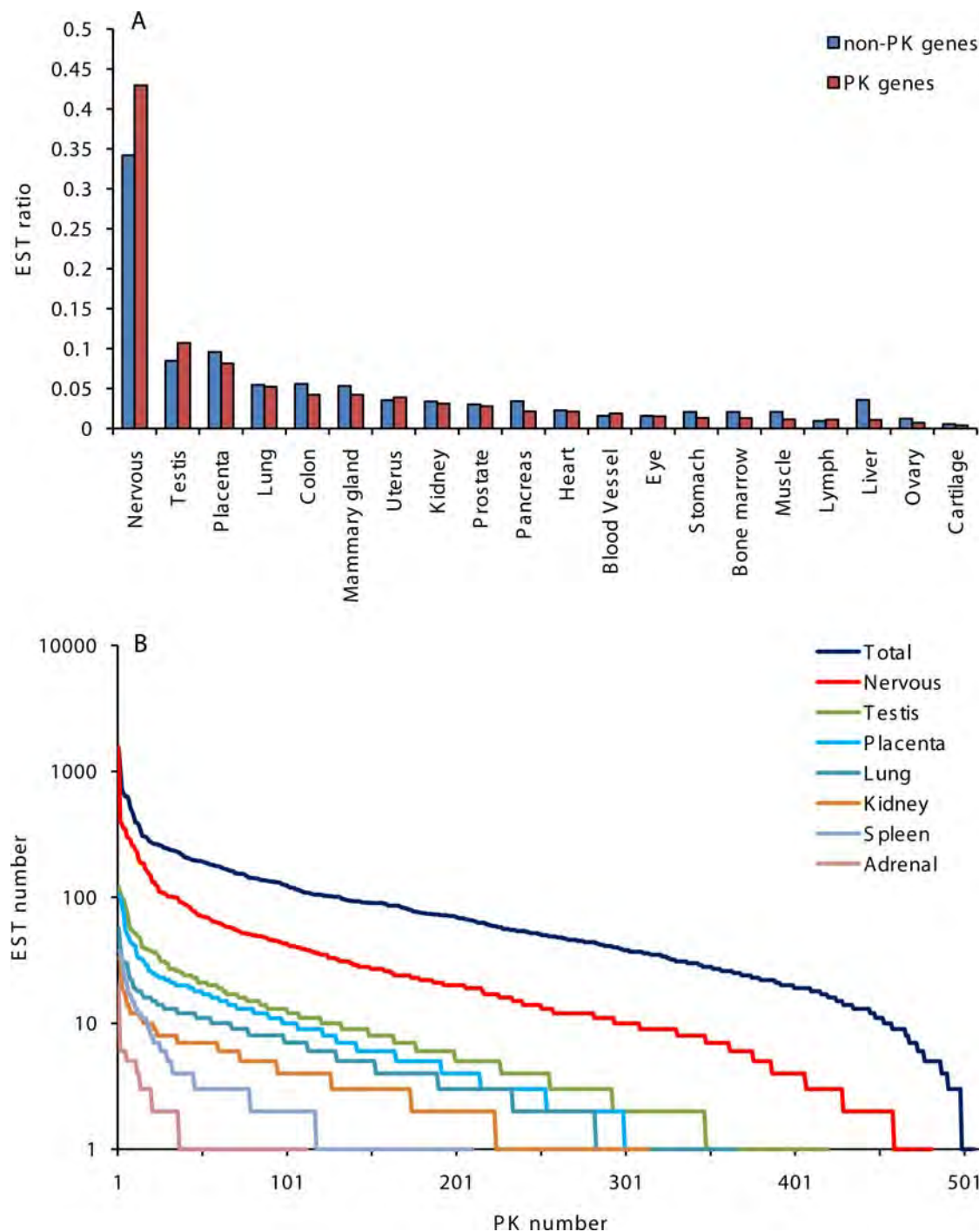


Figure 1. Relative tissue distribution and abundance of EST for 7,711 non-PK genes and 512 PK genes in GenBank, release 162. A. Relative tissue distribution of gene-specific EST for non-PK and PK genes. **B.** Abundance of PK-specific ESTs in libraries from normal human tissues. The data were graphed as EST number versus PK rank. doi:10.1371/journal.pone.0003599.g001

To analyze gene structural features associated with transcription levels, we selected groups of high and low transcribed PK genes. For both groups, we analyzed length, GC-content, and human-mouse sequence conservation in gene functional domains. The proximal 3 kb spacer regions immediately upstream from the translation start site that harbor promoters and the majority of known transcription factor sites in humans were analyzed separately. Results of this analysis are presented in Table 1 and Figure 2. Several structural

features were associated with active transcription and elevated mRNA levels. Primary transcripts and introns of high expression genes were significantly shorter than primary transcripts and introns of low expression PK genes ($p < 0.05$). Consistent with published data [24], this trend was also observed for non-PK genes. High expression PK genes also possessed longer ($p < 0.03$) and a more conserved ($p < 0.02$) 5'UTRs with significantly higher GC-content, and significantly more conserved 3'UTRs ($p < 10^{-4}$) with extended

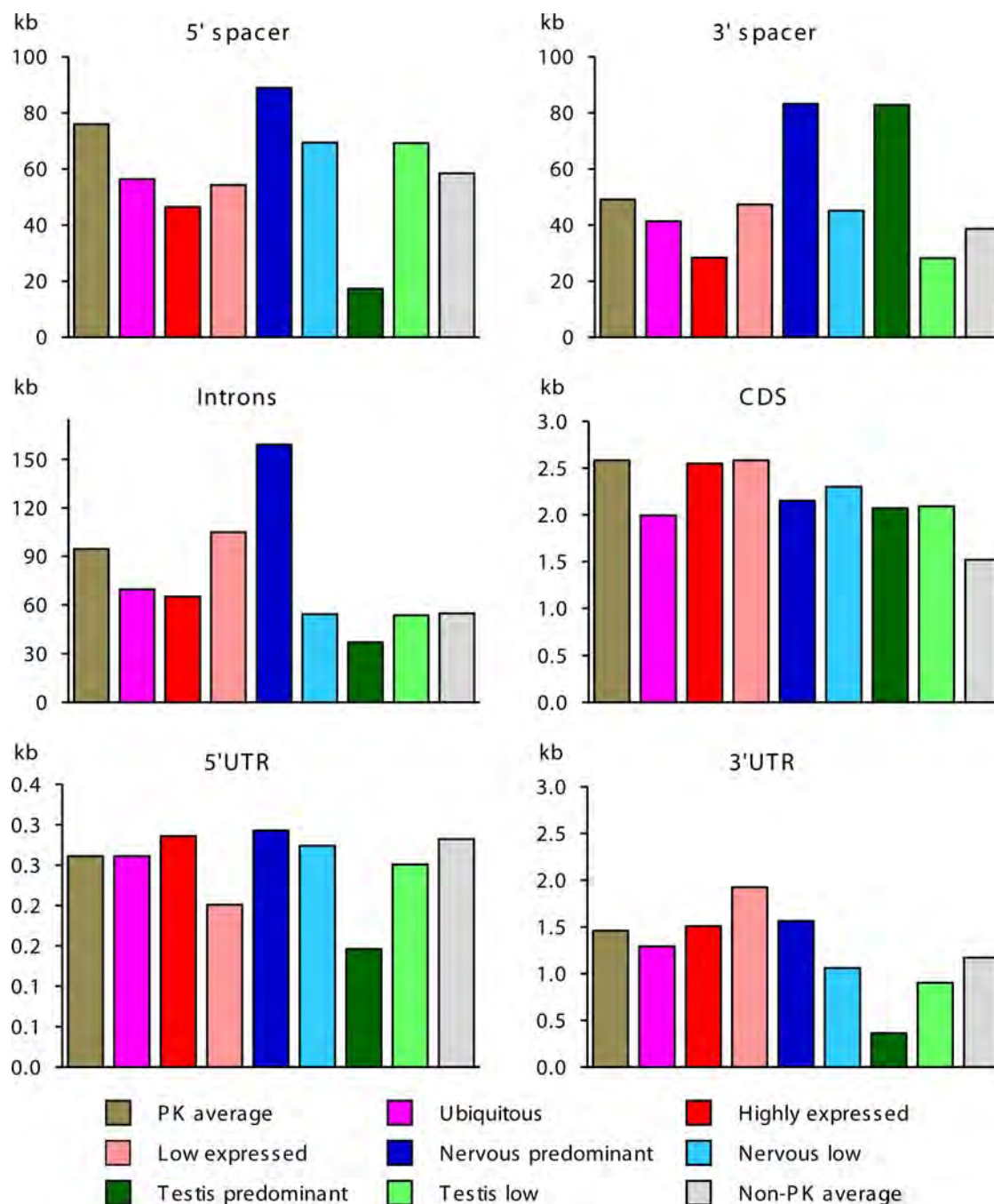


Figure 2. The length of the structural domains in human PK and non-PK genes. The following groups of differentially expressed PK genes were analyzed: all PK genes, high expression genes (75 genes with highest EST numbers), low expression genes (75 genes with lowest EST numbers), ubiquitously expressed genes, genes up-regulated in the nervous or testicular tissue, genes down-regulated in the nervous or testicular tissue. Sequence conservation was evaluated relative to mouse gene orthologs. CDS of extremely large PK titin was excluded from this analysis. Data are presented as averages.

doi:10.1371/journal.pone.0003599.g002

footprints, relative to low expression genes. We found no association between expression levels and the length of mature mRNA, the size of the protein and GC content in distant spacers, introns, and primary transcripts. Similar results were obtained for the mouse PK genes (Figure S1 and data not shown).

We analyzed structural features of PK genes associated with breadth of expression (defined as the number of tissues where a gene is expressed). We observed strong negative correlation between the size of pre-mRNA and the number of expressing

tissues ($R = -0.67$, $p = 1.22 \times 10^{-6}$, Figure 3A). We also found similar correlation for non-PK genes, which was primarily due to smaller size of introns in broadly expressed genes.

Characteristic features of PK genes associated with expression in brain and testis

Distribution of ESTs in tissue libraries (Figure 1A) and expression profiling (Table S1) suggest that the brain and testis

Table 1. GC-content and human-mouse sequence conservation in structural domains of human PK and non-PK genes.

Features	Non- PK genes	PK genes	PK expression level		PK expression breadth				
					Ubiqui-tous	Nervous tissue		Testis	
			High	Low		Up-reg	Down-reg	Up-reg	Down-reg
5'-spacer									
Conservation, %	56.48	55.28	53.62	55.52	55.66	56.59	53.40	58.33	56.60
G+C content, %	45.53	47.39	47.33	47.36	50.65	49.17	49.54	48.05	49.82
Promoter region									
Conservation, %	59.97	60.02	58.26	59.63	61.38	61.17	55.96	60.88	60.37
G+C content, %	48.42	50.83	50.25	49.61	56.27	53.19	52.00	48.52	53.10
Primary transcript									
Conservation, %	59.88	59.03	59.70	59.28	59.17	58.41	57.78	66.54	61.46
G+C content, %	45.08	45.92	46.65	46.99	49.28	46.92	48.27	46.68	49.65
5'UTR									
Conservation, %	72.43	73.54	74.2	70.32	76.61	76.76	68.3	71.66	74.77
G+C content, %	61.27	65.32	66.58	61.01	72.39	64.3	67.12	60.49	67.05
Protein coding regions									
Conservation, %	85.28	87.08	87.81	84.85	88.80	88.57	85.03	84.83	87.40
G+C content, %	51.96	52.20	52.53	52.65	54.37	54.05	53.68	51.23	54.25
Introns									
Conservation, %	56.14	55.03	55.04	54.77	55.05	55.59	53.44	56.10	57.13
G+C content, %	43.64	45.08	45.82	45.96	48.63	46.53	47.40	40.92	48.97
Intron number	9.3	16.3	17	16.8	13.4	18	14	4.5	14.3
3'UTR									
Conservation, %	68.59	69.48	70.83	64.13	71.69	71.19	64.96	66.19	68.47
G+C content, %	42.90	45.20	46.12	46.43	47.39	46.55	46.55	41.44	47.06
3'-spacer									
Conservation, %	58.15	58.59	59.52	58.04	57.69	59.87	56.12	56.82	59.40
G+C content, %	44.11	46.44	47.73	46.77	48.87	47.22	48.34	44.33	46.86

Genomic repetitive elements were excluded from computation of sequence conservation. Data are presented as averages. Gene groups are defined in Figure 2 legend. doi:10.1371/journal.pone.0003599.t001

possess more complex phosphorylation dependent regulatory networks, relative to other organs. To identify gene structural features associated with expression in the nervous and testicular tissue, we analyzed non-overlapping groups of genes predominantly expressed in these tissues. PK genes up-regulated in the brain and testis were compared to control groups of ubiquitously expressed PK genes, and genes down-regulated in these organs. Overall gene organization and features of functional domains significantly differed between these groups (Table 1, Figure 2). Genomic loci and spacer regions of PK genes up-regulated in the nervous tissue were generally longer than those of ubiquitously expressed PKs ($p < 0.0005$) and other analyzed PK groups. Similarly, primary transcripts and introns of PK up-regulated in the nervous tissue were dramatically longer than those of ubiquitously expressed PK genes ($p < 0.0004$ and $p < 0.0005$, correspondingly) and PK genes of other groups.

In contrast, genes up-regulated in the testis were significantly more compact than ubiquitously expressed PK genes ($p < 0.05$) and genes predominantly expressed in nervous tissue ($p < 0.005$), with shorter transcribed regions and smaller number of introns. Testicular PK genes had two to three times shorter 5'-spacers ($p < 0.005$) with significantly lower GC content ($p < 0.02$) in the promoter regions than ubiquitously expressed PKs genes and

genes up-regulated in nervous tissue (Table 1, Figure 2). Testis-specific PK transcripts carried the shortest and least conserved UTRs among all analyzed groups of PK transcripts.

Evolutionary divergence of the human and mouse PK genes

For evaluation of evolutionary divergence, we constructed detailed alignments for human and mouse PK genomic loci. Here we present data for 497 orthologous gene pairs that yielded complete collinear alignments of the transcribed regions, 5'- and 3'-spacer regions, collectively covering over 64 Mb of the human genome. Incomplete alignments that missed spacer regions due to deletions or genomic translocations were not used in our analysis. To compare evolutionary divergence of PK genes with overall trends for other genes, we constructed alignments for control group of 7,711 well annotated orthologous human and mouse non-PK genes.

Protein coding regions of the human and mouse PK orthologs were highly conserved (over 80% identity in nucleotide sequences). To evaluate selection pressure on coding sequences, we calculated levels of non-synonymous (K_a) and synonymous (K_s) human-mouse nucleotide substitutions in the protein coding regions of PK and non-PK genes using Yang's model [25]. Results of these

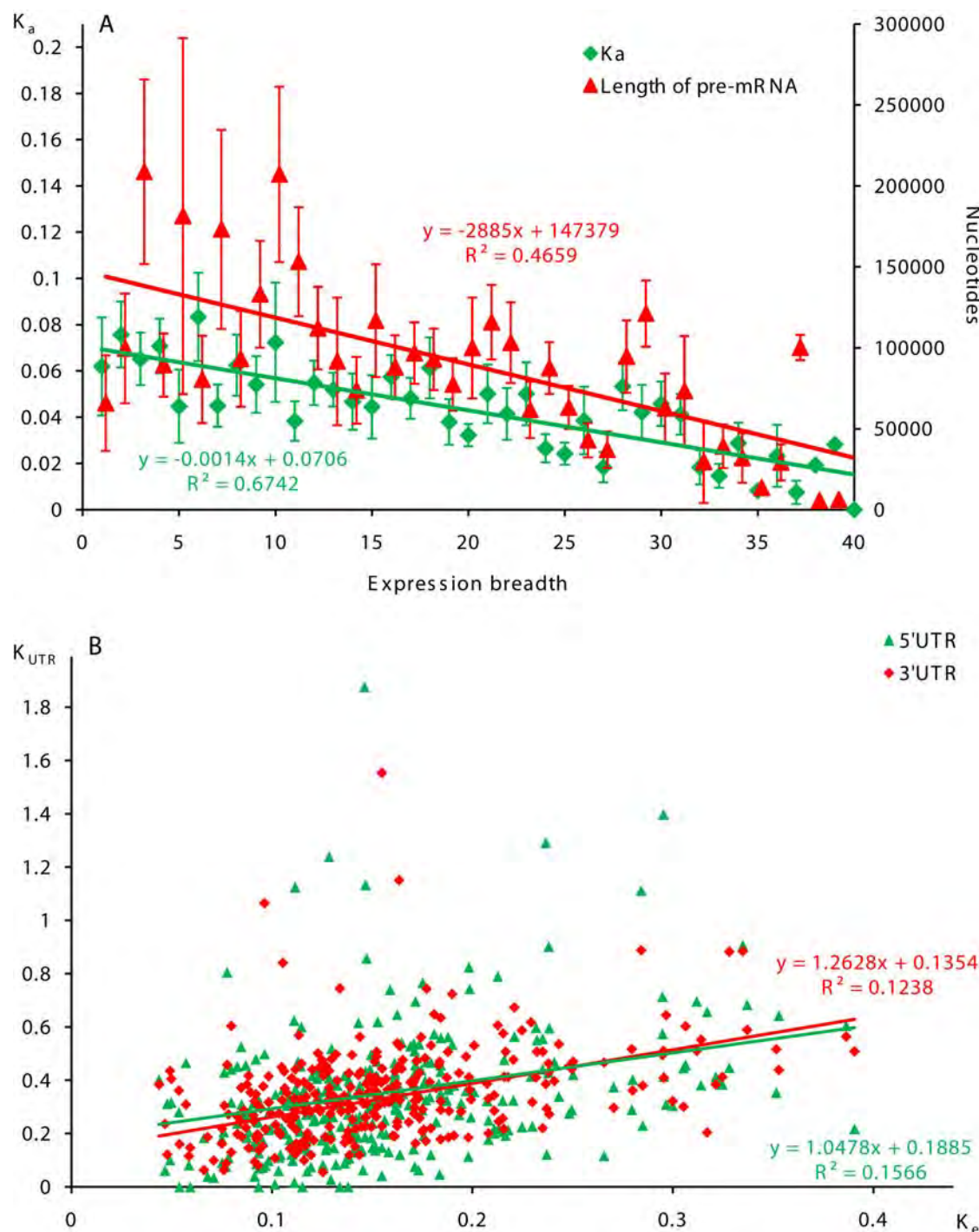


Figure 3. Correlations between the level of human-mouse evolutionarily divergence in protein coding and untranslated regions of human PK genes, expression breadth, and the size of pre-mRNA. A. Correlation between non-synonymous divergence, PK expression breadth and the size of pre-mRNA. Breadth of gene expression was estimated as the number of organ and tissue sources of gene-specific ESTs. Data are presented as averages and SEM. B. Correlations between evolutionarily divergence in protein coding regions and UTRs of PK genes.
doi:10.1371/journal.pone.0003599.g003

calculations are presented in Table 2. The Wilcoxon rank sum test showed that the average K_a values and K_a/K_s ratios were significantly lower for PK coding regions, relative to non-PK genes, indicating stronger purifying selection on PK amino acid sequences. Evolutionary changes in PK protein-coding regions were not homogeneous. Kinase modular domains in CDSs were more conserved than inter-domain regions. Within the group of

PK genes, selective pressure on non-synonymous sites varied significantly depending on expression levels and the number of tissues in which genes were expressed. The level of non-synonymous substitutions in PK genes negatively correlated with breadth of gene expression ($R = -0.82$, $p = 1.73 \times 10^{-9}$, Figure 3A), which is consistent with a general trend for non-PK genes and correlations observed in protein coding regions of non-regulatory

Table 2. Human-mouse evolutionary divergence in the protein coding and untranslated regions of domains of human PK and non-PK genes.

K values	Non- PK genes	PK genes	PK expression level		PK expression breadth				
					Ubiqui-tous	Nervous tissue		Testis	
			High	Low		Up-reg	Down-reg	Up-reg	Down-reg
5'-spacer									
K _{5'}	0.346 (0.002)	0.339 (0.013)	0.322 (0.024)	0.356 (0.034)	0.298 (0.026)	0.288 (0.041)	0.396 (0.039)	0.325 (0.047)	0.340 (0.035)
Protein coding regions									
K _e	0.171 (0.001)	0.159 (0.003)	0.145 (0.006)	0.186 (0.009)	0.135 (0.006)	0.149 (0.008)	0.177 (0.010)	0.189 (0.013)	0.144 (0.011)
K _a	0.073 (0.001)	0.050 (0.002)	0.042 (0.005)	0.060 (0.005)	0.029 (0.003)	0.031 (0.012)	0.057 (0.008)	0.095 (0.012)	0.046 (0.009)
K _s	0.558 (0.002)	0.548 (0.007)	0.540 (0.017)	0.560 (0.018)	0.530 (0.020)	0.538 (0.026)	0.609 (0.025)	0.510 (0.042)	0.540 (0.017)
K _a /K _s	0.128	0.089	0.080	0.110	0.050	0.090	0.090	0.190	0.080
3'UTR									
K _{3'}	0.379 (0.002)	0.376 (0.012)	0.371 (0.038)	0.431 (0.018)	0.322 (0.022)	0.357 (0.028)	0.411 (0.023)	0.447 (0.047)	0.379 (0.021)
Introns									
K _i	0.545 (0.001)	0.564 (0.006)	0.556 (0.013)	0.571 (0.014)	0.579 (0.020)	0.533 (0.013)	0.605 (0.021)	0.571 (0.046)	0.508 (0.016)

Evolutionary divergence in the protein coding regions (K_a), 5'UTRs (K_{5'}), and 3'UTRs (K_{3'}) was calculated using Kimura's two parameter model [28]. Rates of synonymous (K_s) and non-synonymous (K_a) divergence were calculated using Yang's model [25]. Gene groups are defined in Figure 2 legend. Data are presented as averages and the standard error of the mean (SEM, shown in the parentheses).

doi:10.1371/journal.pone.0003599.t002

genes [26,27]. The level of non-synonymous substitutions in PK genes also negatively correlated with gene expression levels ($R = -0.72$, $p = 7.35 \times 10^{-8}$, Figure S2) and positively correlated with the size of pre-mRNA ($R = 0.39$, $p < 0.01$).

On average, protein coding sequences of ubiquitous PKs evolved slower than those of differentially expressed PKs, as seen from their low K_a values (Table 2). PKs with restricted tissue expression (rank of expression breadth ≤ 5) evolved significantly faster ($p < 0.001$) than broadly expressed genes (rank of expression breadth > 30). However, the group of PKs with restricted tissue expression is evolutionarily diverse, as seen from the high values of K_a standard errors for these kinases (Figure 3A), indicating strong variability in rates of their evolution. For example, PKs up-regulated in the highly proliferative testicular tissue evolved almost three times faster than PKs up-regulated in the minimally proliferative nervous tissue which evolved slow, similar to broadly expressed PKs, as seen from their low K_a values. Contrarily, K_s values in the protein coding regions did not differ significantly between ubiquitously and differentially expressed PK groups and did not correlate with gene expression patterns (Table 2), indicating similar levels of synonymous mutations. These results indicate that the differences in K_a values observed between the groups of ubiquitously and differentially expressed PKs are not caused by regional variations in the neutral mutation rate and reflect increased selective pressure on amino acid sequences.

Interestingly, PKs down-regulated in the nervous tissue and PKs with generally low transcription levels also displayed increased divergence in amino acid sequences. Same trends were observed in 5' and 3'UTRs.

We compared evolutionary rates in transcribed domains of PK and non-PK genes by evaluating the human-mouse evolutionary divergence in 5'UTRs (K_{5'}), CDSs (K_e), introns (K_i), and 3'UTRs (K_{3'}) using Kimura's two parameter model [28]. PK genes are characterized with lower K_e values, relative to non-PK genes, reflecting higher constraint on amino acid sequences. We also observed increased K_i values in PK introns, as compared to non-PK introns. As shown in Table 2, evolutionary divergence was

significantly lower for both PK and non-PK genes in 5'UTRs ($p < 10^{-7}$) and 3'UTRs ($p < 10^{-5}$), relative to introns. We found significant positive correlations between levels of evolutionary divergence in CDSs and 3'UTR, in CDSs and 5'UTRs of PK genes (Figure 3B). Similar to K_a values, K_{3'} values inversely correlated with breadth of gene expression ($R = -0.11$, $p < 0.01$). Positive correlation between K_e and K_{3'} values was observed for PKs predominantly expressed in nervous tissue, and for other differentially expressed PK groups (Figure 4). This trend was also observed for slow evolving ubiquitous PKs.

Regulatory signals in PK genes associated with transcription levels

Taking into consideration strong relationships between gene transcription levels, evolutionary conservation, and the structure of regulatory domains, we attempted to identify evolutionary conserved DNA elements that regulate gene expression. For regulatory elements associated with transcript abundance, we searched for motifs over-represented in conserved promoter regions of high expression PK genes using the discriminating matrix emulator (DME) program. Conserved promoter sequences of low expression PK genes were used as a background set in this analysis. DME search revealed a number of motifs ranging from 6 to 10 nt which were significantly over-represented in promoters of high expression PK genes. Some of the characteristic over-represented motifs are shown on Figure S3, and the top 50 over-represented motifs are presented in Table S1. Promoters of high expression, actively transcribed PK genes were enriched with GC-rich motifs. In contrast, promoter regions of low expression PK genes were dominated by AT-rich low complexity motifs (data not shown). To identify potential binding sites for transcription factors, we searched through TRANSFAC database for position frequency matrices that match the motifs found by DME. Most common motifs were identified as conserved and degenerate binding sites for transcription stimulating proteins Sp1 and Sp3, core binding sequences for activator protein AP-2, characteristic for increased promoter performance, and other transcription factors. Some of

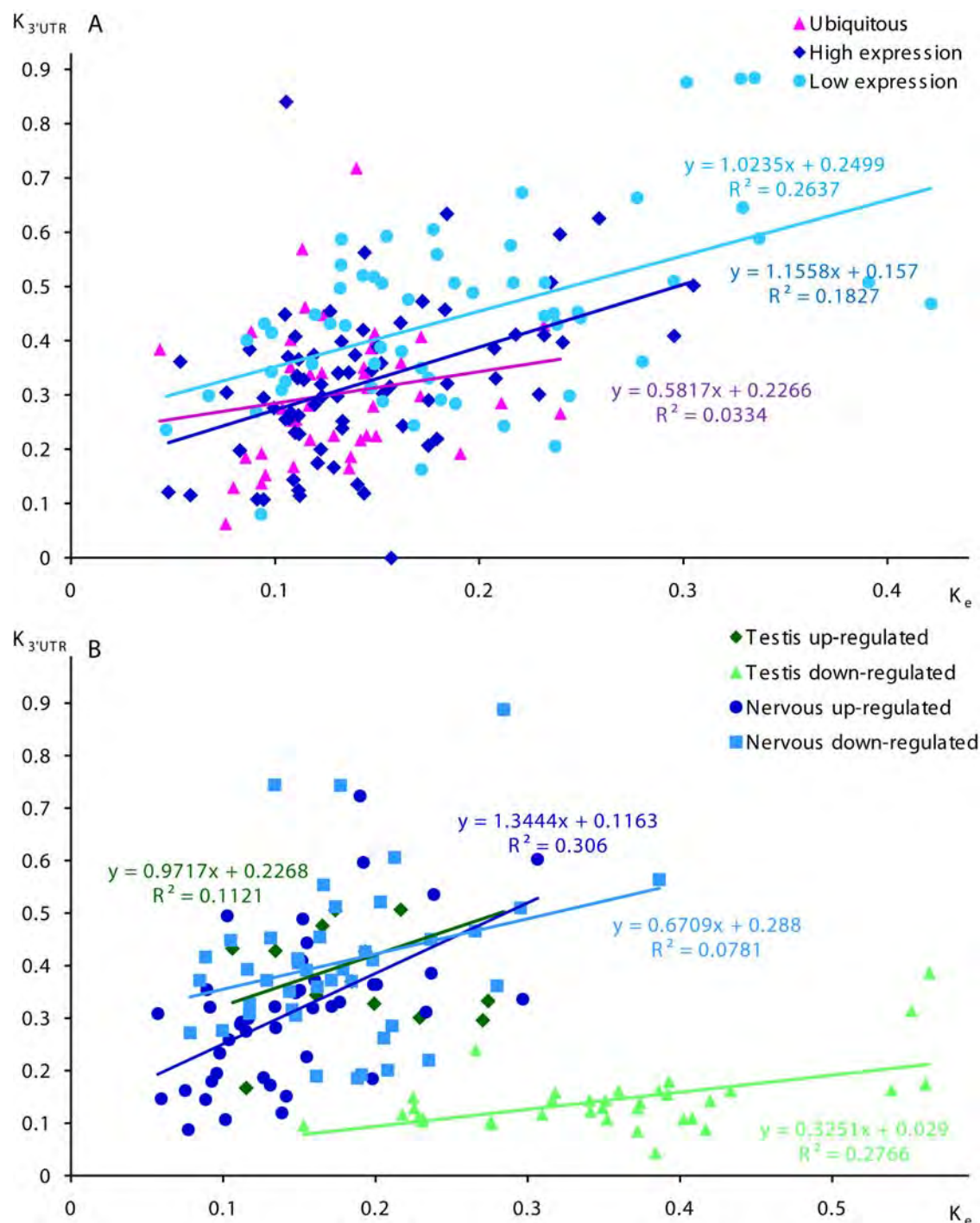


Figure 4. Correlations between the level of human-mouse evolutionary divergence in protein coding regions and 3'UTRs of ubiquitous and differentially expressed human PK genes. **A.** Correlations between evolutionarily divergence in protein coding regions and 3'UTRs of high, low, and ubiquitously expression PK genes. **B.** Correlations between evolutionarily divergence in protein coding regions and 3'UTRs of PK genes up-regulated and down-regulated in nervous and testicular tissues.
doi:10.1371/journal.pone.0003599.g004

the over-represented motifs were not identified as recognition sites for known DNA-binding proteins.

To identify sites in 5'UTRs associated with transcript abundance, we performed a search for evolutionarily conserved over-represented sequence elements using the SiteBD program. 5'UTRs of abundant transcripts have significantly higher GC-content and were enriched three-fold with repetitive GGCGGCCGC motifs

($p < 5 \times 10^{-55}$), complementary CCGCCGCCG motifs ($p < 9 \times 10^{-39}$), and other GC-containing sites, as compared to rare transcripts (Table S1). Several motif variants differing by nucleotide shifts were found. Unlike typical transcription factor binding sites, these motifs possessed a low degree of degeneration, mostly resided in 5'UTRs and were rarely encountered in the proximal spacers regions.

To identify sites of potential interaction with ribosomes, we evaluated hybridization affinity of abundant and rare PK transcripts to 18S ribosomal RNA. As seen from Figure 5A, 5'UTRs of abundant transcripts possessed two to three-fold higher potential to form intermolecular duplexes with 18S ribosomal RNA, relative to 5'UTRs of rare PK transcripts. This effect was observed for theoretically predicted 18S rRNA “clinger” sites

(data not shown), and also for an experimentally confirmed “clinger” element [29], which base pairs to a core of the translation enhancer commonly occurring in the 5'UTR (Figure 5B).

It was shown earlier that selection may be operating in the protein coding regions on most variable synonymous positions to maintain a more stable and ordered mRNA secondary structures

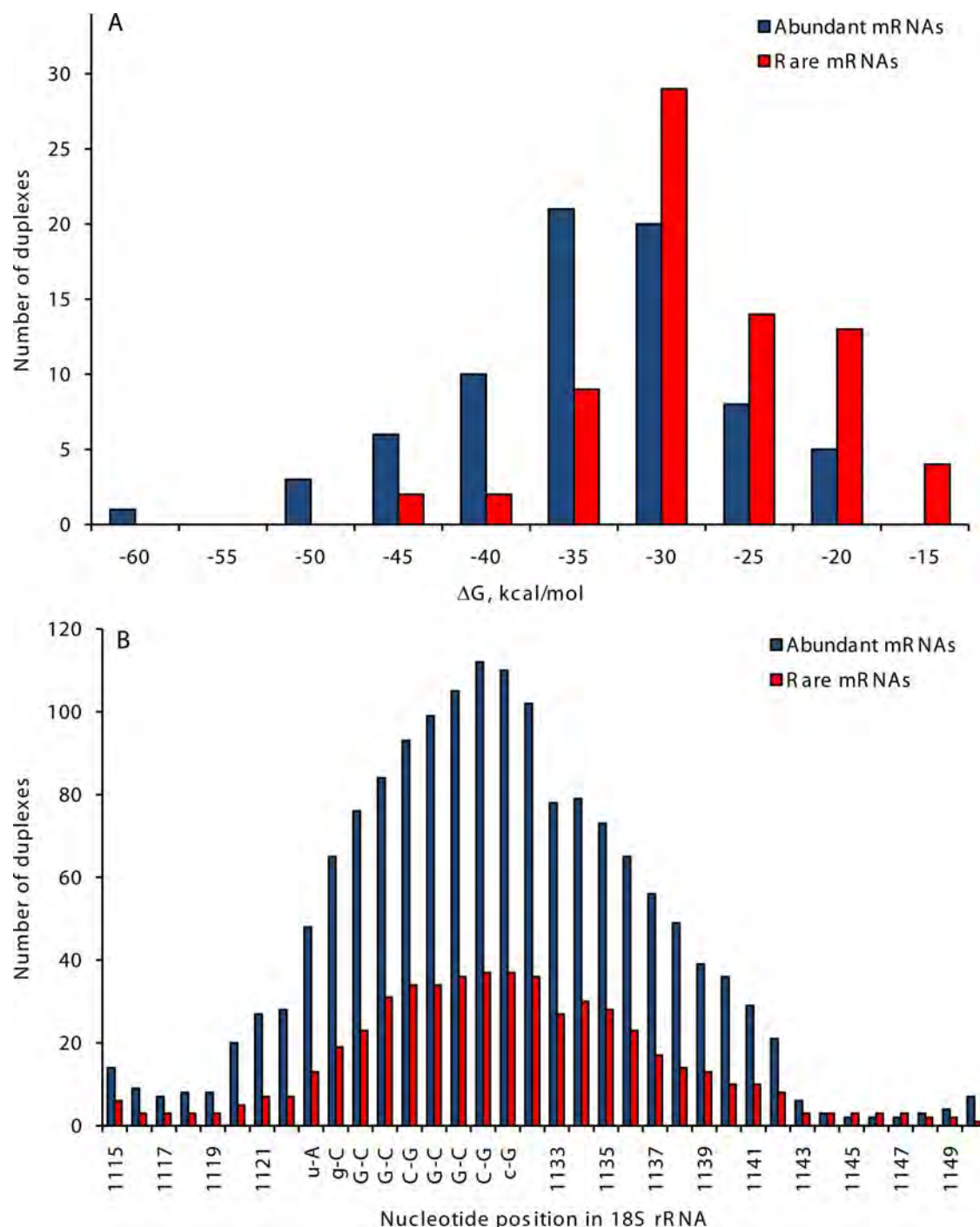


Figure 5. Hybridization affinity of 5'UTRs of PK genes to human 18S ribosomal RNA. A. Distribution of calculated energy of predicted duplex formation (ΔG) between human 18S rRNA and 5'UTRs of abundant and rare PK transcripts. Site of duplex formation between experimentally confirmed 18S rRNA “clinger” element and a core of 5'UTR translation enhancer [29] is shown. **B.** Hybridization affinity of abundant and rare PK transcripts to experimentally verified 18S rRNA “clinger” element [29]. doi:10.1371/journal.pone.0003599.g005

[30]. To evaluate secondary structures in 5'UTRs, we computationally “folded” sequences of mature PK transcripts. In agreement with the results of transcriptome-wide analysis of mammalian mRNA folding, we found that secondary structures are frequently formed in PK transcripts between UTRs and CDSs in the vicinity of the start and stop codons (Figure S4). Sequences immediately upstream from the start codon have strong hybridization affinity to the N-terminal coding sequences, thus

promoting formation of local hairpin structures where the start codon is positioned at the end of a hairpin in a relaxed loop (Figure 6). This type of secondary structure was facilitated by the enhanced GC-content in the N-terminal protein coding regions, and was observed in both abundant and rare transcripts. However, thermodynamic stability of these characteristic hairpin structures was significantly higher for abundant PK transcripts, than for rare transcripts ($p < 10^{-5}$).

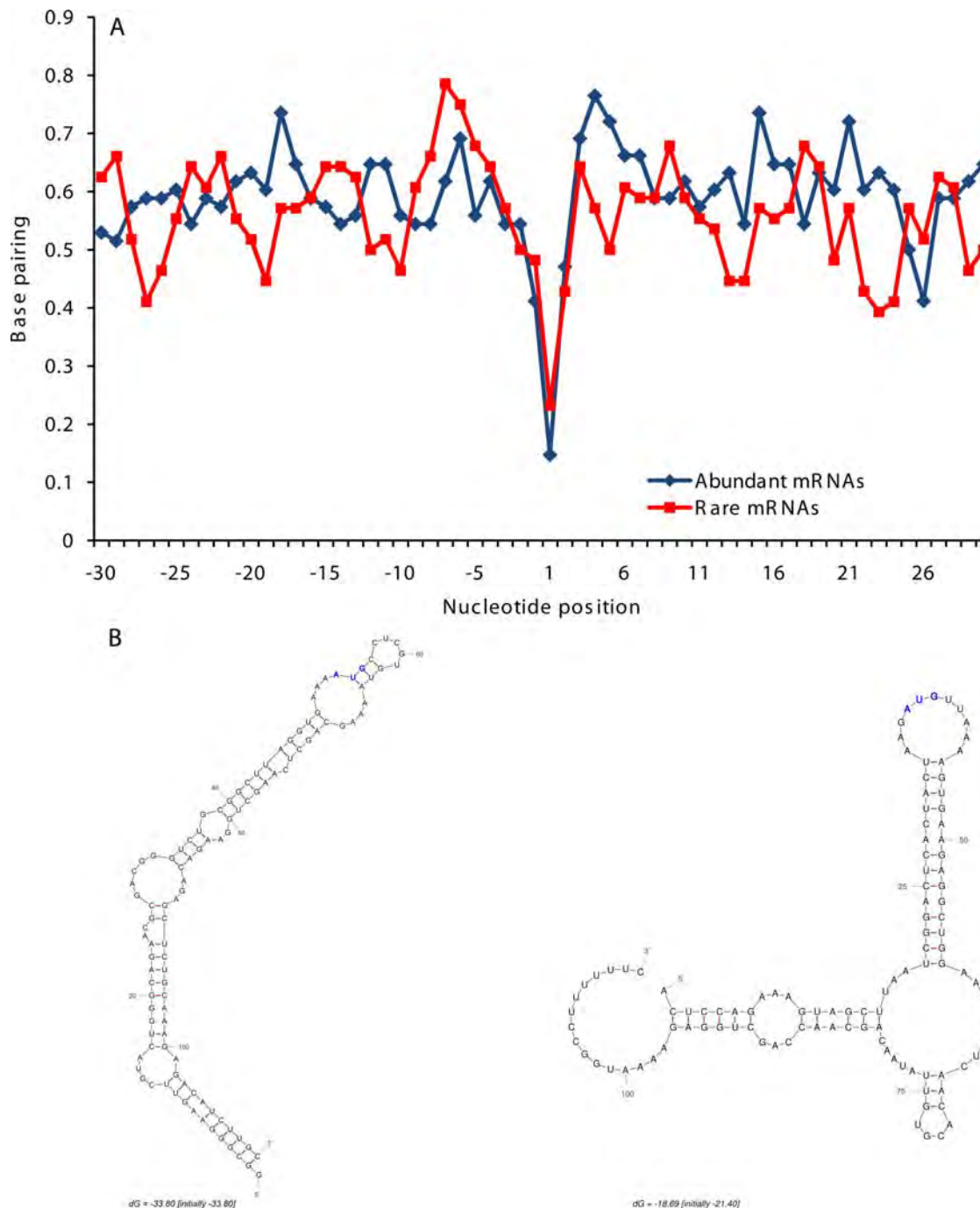


Figure 6. RNA secondary structures in the vicinity of the start codon. A. Profiles of nucleotide base pairing in the vicinity of the start codon for abundant and rare PK transcripts. Nucleotide positions are shown relative to translation start codon. **B.** Examples of predicted local secondary structures in the vicinity of the start codon for an abundant PK transcript (VRK1, NM_003384) and a rare PK transcript (MYLK4, NM_001012418). Start codons are shown in blue.

doi:10.1371/journal.pone.0003599.g006

Abundant PK transcripts also carry significantly more conserved 3'UTRs, relative to rare transcripts ($p < 10^{-4}$). No significant difference in nucleotide levels was observed between 3'UTR of abundant and rare PK transcripts, which had uniformly high AT content (Table 1).

Nervous tissue-specific regulatory signals

PK genes up-regulated in the nervous tissue had significantly longer 5'-spacers and introns than genes down-regulated in the nervous tissue. These extended gene loci may harbor binding sites for nervous tissue-specific transcription factors. To identify brain-

specific regulatory elements, we analyzed conserved syntenic regions of PK genes predominantly expressed in the nervous tissue with the DiRE program. Whole gene loci, including promoters, UTRs, introns, and distant intergenic and spacer regions were included in this analysis. Conserved syntenic regions of PK genes with similar overall expression levels and low expression in the brain tissue were used as a background set. Typical transcription factor binding sites overrepresented in evolutionarily conserved brain-specific PK genes are shown in Figure 7, and the top 30 overrepresented motifs are presented in Table S2. As seen from the Table, binding sites for POU, Pit, Pbx,

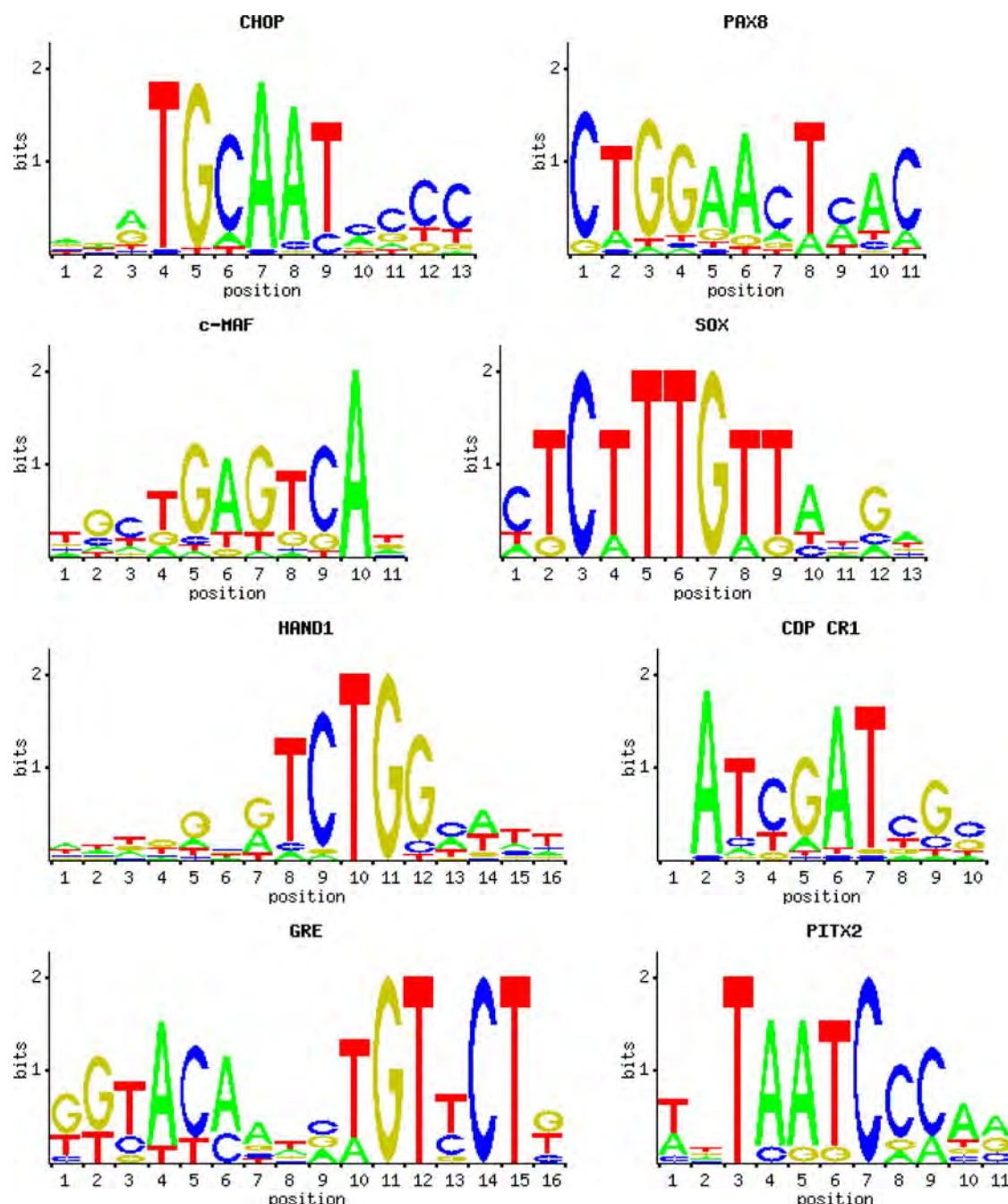


Figure 7. Characteristic transcription factor binding sites over-represented in evolutionarily conserved regions of PK genes predominantly expressed in nervous tissue.

doi:10.1371/journal.pone.0003599.g007

Pax, Olf, Meis and other neuron-specific transcription factors that perform specific functions in the central nervous system were highly overrepresented in evolutionarily conserved regions of PK genes predominantly expressed in the nervous tissue.

We searched for overrepresented conserved motifs in promoters of PK genes up-regulated in the nervous tissue using the DME program. Conserved promoter sequences of genes down-regulated in the nervous tissue were used as background in this analysis. Promoter regions of PK genes predominantly expressed in nervous tissue contained numerous over-represented sites that compositionally and textually differed from the sites associated with transcript abundance (data not shown). They were enriched with CTGG, TGCA, TCTGG, CAATC and CTGA motifs that constituted nucleotide core sequences for neuron-specific transcription factors identified with the DiRE program.

The number of predicted functional signals in 3'UTRs correlated with sequence conservation, indicating a significant level of evolutionary conserved posttranscriptional regulation in nervous tissue. To evaluate potential regulation of PK expression by RNA inhibition, we analyzed hybridization affinity of annotated human miRNAs to 3'UTRs of human PKs up-regulated and down-regulated in nervous tissue. Remarkably, we observed a significant difference in the number of binding sites for neuron-specific miRNAs between the two groups of PK transcripts (Figure 8). Transcripts rarely encountered in the nervous tissue were enriched 2–3 fold with binding sites for neuron-specific miRNAs, which likely facilitated targeted degradation of these transcripts in the nervous tissue through the RNA inhibition mechanism.

Discussion

Gene architecture and expression levels

Physiological complexity of an organism is largely dependent on the regulation of gene expression and correlates with the length of non-coding DNA, and the number of cis-control elements in genome [31]. It is estimated that as much as a third of the human genome controls chromosome replication, condensation, pairing, segregation, and gene expression [32]. The regulatory array of a typical mammalian gene locus consists of a core promoter, and proximal promoter elements (located within ~3 kb in the

upstream spacer region, 5'UTR and sometimes within the first exon of a gene), and distant enhancers, silencers, chromatin insulators, and scaffold/matrix attachment regions that can be scattered within 100 kb or more from the transcription start site. The size of a gene locus inversely correlates with GC content and depends on gene location within the complex landscape of the human genome, which is dominated by extensive GC-rich regions (isochores) alternating with GC-poor isochores. GC-rich isochores are densely populated with compact genes, while GC-poor isochores are low gene density regions populated with larger genes [33]. PK genes occupy larger genomic loci than other mammalian genes. Our results indicate that human and mouse PK genes expressed at different levels and in different tissues have similar GC-content in intergenic regions, suggesting that they are not confined to GC-rich isochores. Noteworthy, 5'-spacer regions and introns of the human PK genes were significantly (~15%) longer than the corresponding regions of the mouse genes, implying a more complex regulation of transcription.

There is a strong negative correlation between gene length and the level of gene expression. Highly expressed human housekeeping genes are generally shorter and more compact than genes that are expressed at lower levels, which has been explained by selection pressure to reduce metabolic costs of transcription [24,34]. Our results (Figure 2), consistent with published data for non-regulatory genes [34,35] indicate that PK genes have significantly longer spacer regions, introns, primary transcripts, and encode larger and more complex proteins than average housekeeping and non-housekeeping mammalian genes. Within the PK group, we observed major differences in gene architecture that are related to gene expression levels and tissue-specific expression. Gene transcription levels and breadth of expression negatively correlate with the length of introns and the size of primary PK transcripts. Primary transcripts of highly and ubiquitously expressed genes are significantly shorter than primary transcripts of low expression PK genes. Average length of protein coding regions in ubiquitously expressed PK genes tends to be shorter than in other groups, which may reflect evolutionary pressure to reduce costs of translation for housekeeping protein kinases.

Transcription efficiency of a gene is determined by many factors, most important of which are the structure of promoter, the

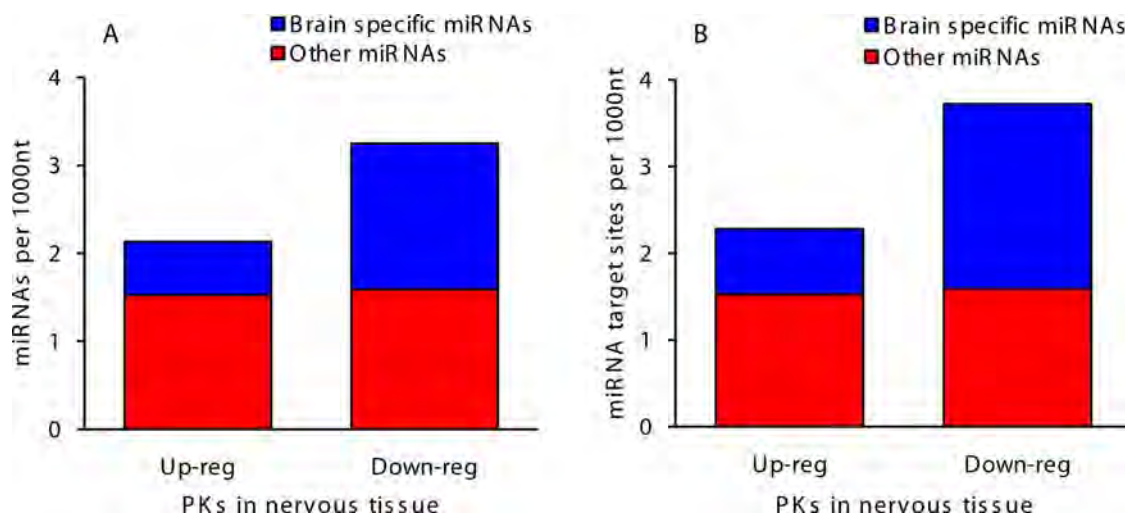


Figure 8. Hybridization affinity of human miRNAs to 3'UTRs of human PK genes up-regulated and down-regulated in nervous tissue. A. Predicted numbers of miRNAs hybridizing to 3'UTRs. B. Predicted numbers of miRNA target sites in 3'UTRs.

doi:10.1371/journal.pone.0003599.g008

array of transcription factor binding sites, and the length of transcribed units. Mammalian transcription factor binding sites are usually organized in clusters that contain several phylogenetically conserved sites for a few different transcription factors [36]. Degenerate sites are enriched around the cognate binding sites in orthologous mammalian promoters [37], which may facilitate recruitment of transcription factors to DNA regulatory regions and a robust transcription response (for discussion of possible molecular mechanisms, see [37]). Our data show that GC content in promoter regions and 5'UTRs of ubiquitously expressed PK genes is stably higher, relative to other groups.

Proximal 5'-spacers of actively and ubiquitously transcribed PK genes carry complex cis-regulatory modules enriched with conserved and degenerate GC-rich binding sites for transcriptional activators of Sp, AP, and other families. Abundance of these sites is indicative of increased promoter strength and correlates with gene transcription levels.

PK expression in nervous and testicular tissue

Human brain and testes employ large groups of tissue-specific PKs (Table S1), implying increased complexity of phosphorylation-dependent regulatory network that controls the functioning and homeostasis of these organs. We observed major differences in gene architecture and expression between PK groups up-regulated in nervous and testicular tissues. Adult brain is characterized with significant transcriptional and very low mitotic activity. Our data demonstrate that PK genes up-regulated in nervous tissue possess long transcribed regions with numerous extended introns, an indication of increased complexity of transcription regulation. They also carry extended spacer regions which may be required for accommodation of distant neuron-specific enhancers and extended 5'UTRs, an indication of increased post-transcriptional regulation.

In contrast, testis-specific PK genes are compact, with short and low conserved regulatory regions. These properties of testis-specific PK genes likely reflect simpler transcriptional control by testis-specific transcription factors that turn transcription on and off at specific stages during linear progression of germ cell development. Male germ cell development and differentiation require massive production of testis-specific transcripts and isoforms. Transcription during germ cell development peaks at meiotic and post-meiotic phases [38]. This increased expression fulfils the requirement for the rapid accumulation of large amount of transcripts in haploid round spermatids in preparation for radical restructuring of the cytoplasm and chromatin condensation at the following stage of elongated spermatids. After this final burst, transcription ceases in elongated spermatids, chromosomes are stripped of nucleosomes and densely packed in sperm heads. We speculate that general compactness of genomic loci of PK genes predominantly expressed in the testis [39], small size of their pre-mRNAs and reduced number of introns (Table 1) are likely dictated by the need for intense transcription in germ cells during the relatively short developmental time frame. Interestingly, several testicular PKs (TSSK1, TSSK2, SSTK and Haspin) are evolutionarily young, intronless, and supposedly originated by retrotransposition [7]. Same tendency for gene compactness was observed for proliferatively active placental tissue (data not shown).

Evolutionary divergence of PK genes

Protein coding sequences of PK genes are evolutionarily more conserved than coding sequences of other genes (Table 2), which is likely due to tighter purifying selection on catalytic kinase domains. Evaluation of sequence divergence between the human and mouse PKs revealed differing rates of evolution for different gene groups.

Ubiquitous PK genes evolve slower, relative to differentially expressed genes. Elevated conservation in ubiquitous PKs likely reflects increased selection pressure on housekeeping kinases. These data are in agreement with the observation that proteins with a broad range of tissue expression tend to be more conserved than those expressed in one or few cell types [26].

Our results demonstrate that genes preferentially expressed in different tissues evolve with differing rates. Rates of evolutionary divergence of PK genes up-regulated in nervous and testicular tissues correlated with the proliferative activity of the tissue and with the length of transcribed gene domains. Consistent with the published results [7], the group of testis-specific kinases is the most divergent between the human and mouse in the protein coding regions. This divergence is reflective of higher evolutionary rates of testicular PKs. Conversely, PK genes preferentially expressed in nervous tissue are more conserved between the human and mouse in the coding regions and UTRs, indicating elevated selection pressure on amino acid sequences and on RNA regulatory sequences.

Interestingly, our results demonstrate increased evolutionary divergence for the groups of PK genes with generally low expression levels and genes with low expression in the nervous tissue, relatively to ubiquitous PK genes. A likely explanation is that the low expression group contains genes selectively expressed in low abundant cell types. This group may also contain evolutionarily young genes with low expression levels and emerging function. Many of the genes from the second group are preferentially expressed in two or more tissues, suggesting possible diversification or specialization of function, which may be accompanied with evolutionary divergence. It is possible that some of the observed differences in sequence conservation between the human and mouse may be due to the unique physiology of the mouse. Consistent with published reports [40,41], evolutionary conservation in the protein coding regions strongly correlated with conservation in 3'UTRs and 5'UTRs for all gene groups. These correlations were more pronounced for differentially expressed PK genes than for ubiquitously expressed genes. Our results for PK genes are in good agreement with published data for other genes [26] and support the idea that expression patterns affect selection intensity but not mutation rate.

Regulation of expression by 3'UTRs

In eukaryotic cells, transcripts exist as complexes with associated proteins that are essential for mRNA transport across nuclear membrane, stability, and translation. Messenger RNA degradation and translation are tightly coupled events, and efficiency of gene expression is largely dependent on post-transcriptional stability of mRNA. Both mRNA stability and translation efficiency depend on the 3' poly(A) tail which interacts with poly(A) binding protein (PABP). PABP is involved in mRNA circularization by binding together the 5' and 3' ends of mRNA, and also plays a role in translation initiation and mRNA degradation. The major pathway of eukaryotic mRNA decay is initiated with degradation of the 3' poly(A) tail and the loss of PABP, linearization of transcript, and cleavage of the methylated 5' cap structure, followed by 5' to 3' exonucleolytic degradation of mRNA (reviewed in [42]). Other RNA-binding proteins regulating post-transcriptional mRNA stability and turnover specifically interact with AU- and CU-rich sites in 3'UTRs [43]. For example, stability of the epidermal growth factor (EGF) receptor tyrosine kinase mRNA is mediated by two RNA-binding proteins that bind to AU-rich elements in the 3'UTR. The binding affinity of these proteins is down-regulated by the kinase ligand, EGF [44].

Another major regulatory pathway is RNA inhibition. 3'UTRs are recognized and targeted by small non-coding miRNAs that inhibit translation, promote transcript degradation, and often act as tissue-specific regulators of transcript stability [45]. A large set of housekeeping genes involved in basic cellular processes avoid regulation by miRNAs due to short 3'UTRs that are depleted of miRNA binding sites [46]. Our results are suggestive that PK genes are extensively regulated through RNA inhibition in a tissue-specific manner. 3'UTRs of transcripts rarely encountered in the nervous tissue were enriched with binding sites for neuron-specific miRNAs (Figure 8), which likely facilitate targeted degradation of these transcripts in the nervous tissues through the RNAi mechanism. Comparison of our results with published data [47] reveal that sites over-represented in the brain-specific PK transcripts include few ubiquitous miRNA binding sites that are commonly found in human transcripts, suggesting regulation by unidentified miRNAs. 3'UTRs of PK genes predominantly expressed in the nervous tissue were significantly longer and more conserved, as compared to genes expressed in the nervous tissue at low levels. These conserved GC-rich sites participate in the formation of local secondary structures in 3'UTRs which increase the compactness of transcript folding (data not shown) and may be involved in regulation of mRNA stability.

Regulation of expression by 5'UTRs

Recent genetic studies demonstrated that mutations and single nucleotide polymorphism in 5'UTRs affect transcription efficiency, mRNA levels, and have implications in human disease [48,49,50,51]. The length of the 5'UTR negatively correlates with mRNA and protein expression levels. Transcripts of highly expressed housekeeping genes carry short 5'UTRs devoid of strong secondary structures. Conversely, transcripts of low expression regulatory genes controlling cell proliferation, survival and apoptosis carry long 5'UTRs with stable secondary structures and upstream translation start codons [35,52,53]. Similar to other regulatory genes, PK genes possess long and complex 5'UTRs. Surprisingly, our analysis showed that abundant PK transcripts carry longer 5'UTRs with higher GC-content that form more stable secondary structures, as compared to rare PK transcripts. GC-rich elements in 5'UTRs are mostly confined to evolutionarily conserved sequences and could be maintained by selective pressure due to conserved biological function. Our observations suggest that GC-rich elements in 5'UTRs may function at the mRNA level rather than at the DNA level. Increased GC content in 5'UTRs of abundant transcripts allows formation of more stable RNA secondary structures that may serve as scaffolds for RNA-binding proteins, promote a more compact folding and increased mRNA stability.

The majority of translational control events occur at the level of initiation, implicating the 5'UTR as the major site of translational regulation. The cap-dependent initiation of translation is affected by mutations in the 5'UTR and severely hampered by stable secondary structures that can stall the ribosome and inhibit translation [54,55]. Translation of mRNAs encoding regulatory proteins is often initiated via internal ribosome entry or other yet unknown mechanisms [56]. Transcript folding in the vicinity of the start codon favors formation of characteristic local structures where the start codon is positioned at the end of a hairpin in a relaxed loop [30]. This type of secondary structure is preferred in GC-rich 5'UTRs of abundant PK transcripts (Figure 6A) and probably represents adaptation for a more efficient translation of abundant mRNAs. Positioning of the AUG codon at the end of stem-loop may provide a more productive recognition context during cap-independent initiation of translation, when the

ribosome binds directly to internal entry site without unwinding and scanning upstream regions of mRNA molecule.

A role was proposed for 5'UTRs in regulation of translation through intermolecular base-pairing interaction with 18S ribosomal RNA. It was hypothesized [57] and experimentally confirmed [58] that accessible "clinger" regions of 18S rRNA may function as low-specificity mRNA-binding sites allowing a more efficient transcript interaction with the ribosome. This interaction may affect translational efficiency of different subsets of mRNAs. Our data suggest that 5'UTRs of abundant PK transcripts have significantly higher hybridization affinity to 18S rRNA, than 5'UTRs of rare PK transcripts (Figure 5). Most common elements in 5'UTRs of abundant PK transcripts are short GGC repeats. CGGCGG element was recently identified as a core of a translation enhancer commonly occurring in 5'UTRs of mammalian mRNAs, which base pairs to a "clinger" site on 18S ribosomal RNA and facilitates translation initiation [29]. Studies of the functional role of CGG repeats in the 5'UTR of the human FMR1 gene demonstrated that these repeats may exert both positive and negative effects on the efficiency of translation of FMR1 mRNA, depending on repeat length. Long repeats in 5'UTR of FMR1 suppressed translation. However, the presence of short repeats increased translation efficiency in the absence of any change in mRNA levels [59]. Our results are consistent with these data and provide additional support for the ribosome filter hypothesis [58]. Other evolutionarily conserved GC-rich motifs in 5'UTRs of abundant PK transcripts (Table S1) may affect translation initiation through similar mechanisms.

Conclusions

Genomic organization of the human PK superfamily and the structure of gene functional domains significantly differ from other protein coding genes. Our results demonstrate that gene expression levels, expression breadth, and requirements for tissue-specific regulation correlate with genomic architecture. These factors also may contribute to selection pressure in the protein-coding and non-coding DNA regions. Transcription levels and breadth of expression negatively correlate with the length of introns and the size of primary transcript, which is likely due to the necessity to minimize metabolic costs of transcription for abundant mRNAs. It is generally accepted that mammalian ubiquitously expressed genes evolve slower than tissue-specific genes. Here we show that genes up-regulated in different tissues evolve with different rates, and that evolutionary rates correlate with the proliferative activity of expressing tissue and with gene architecture. The observed negative correlation between the length of transcribed gene domains and the proliferative activity of tissues may reflect metabolic constraints and requirements for tissue-specific expression. Our data provide evidence that evolutionarily conserved phylogenetic footprints and structural elements in messenger RNA play roles in regulation of transcript abundance, tissue-specific expression, and translation. All these mechanisms may contribute to the multi-level regulation of PK expression, providing precision control of the key components of the cell signaling pathways that determine cell function and destiny.

Methods

Gene sequences and alignments

Protein kinase names and classification in this paper are presented according to Manning et al [5]. Sequences of the full-length human PK mRNAs were downloaded from http://kinase.com/kinbase/FastaFiles/Human_kinase_rna.fasta and aligned to the sequence of the human genome (<ftp://hgdownload.cse.ucsc>

edu/goldenPath/hg18/chromosomes/), March 2006 assembly. To compare our results for PK genes with overall trends for other genes, we compiled control group of 7,711 randomly selected non-PK genes well-annotated orthologous human and mouse genes that yielded high quality genome alignments. Gene coordinates were downloaded from <http://genome.brc.mcw.edu/cgi-bin/hgTables>. For each human mRNA, a mouse orthologue was found as a best Blast hit with complete mouse mRNA sequences. Only full-length transcripts with links to the RefSeq database were used (<http://www.ncbi.nlm.nih.gov/RefSeq/index.html>). Mouse genome sequences (February 2006 assembly) were downloaded from <ftp://hgdownload.cse.ucsc.edu/goldenPath/mm8/chromosomes/>. Genome coordinates of extended human gene loci were transferred to the mouse genome sequence with UCSC Lift Genome Annotations tool (<http://genome.brc.mcw.edu/cgi-bin/hgLiftOver>). Mammalian genomic repeats were masked and extended genomic loci of orthologous human-mouse genes were aligned with the OWEN program [60] and annotated. In case of alternatively spliced forms, the longest CDSs and UTRs were considered. For the protein coding regions, the alignment of nucleotide sequences was guided by the amino acid sequence alignment. 5' and 3' intergenic regions were considered separately and orientations of gene loci were assigned from the 5' end to the 3' end of a gene. For 98% of PK genes, UTRs were annotated and considered separately. When UTRs are included in the intergenic regions, they usually constitute only a minor fraction of the sequences and can not affect significantly the results of computer analysis of intergenic regions. We additionally analyzed proximal 5'-spacer regions (3 kb upstream from the transcription start site) that contain core promoter and proximal promoter elements. Overall level of similarity was calculated from percent identity in global alignments and sequence lengths. Core hits with ϵ -value lower than 10^{-3} produced by OWEN program were extracted for analysis. Phylogenetic footprints were identified using conservative parameters (match = 1, mismatch = -1, gap = -5/-1) where the final similarity of the extended core regions was higher than 50% and boundaries were matches. The lowest significantly non-random level of similarity for two-sequence alignments (A ~30%, T ~30%, G ~20%, C ~20%) is 42% [61]. Statistical analyses of phylogenetic footprints were conducted using Excel (Microsoft, USA) and our statistical tools [62].

Evaluation of gene expression levels

We evaluated relative transcript abundance using the numbers of gene-specific expressed sequence tag (EST) sequences in GenBank. We used EST approach because it allows a more reliable identification of the transcript identity than microarray data and has a greater potential for quantitative analysis, since EST clone frequency in a library is generally proportional to the corresponding gene expression levels. This approach gives a reasonably accurate approximation of gene expression and was successfully used for studying gene transcription levels and tissue-specific gene expression (for example, [24,26,63,64,65]). We aligned sequences of PK mRNAs with PK-specific ESTs from the human normal tissue EST libraries from GenBank using the program BLAST [66]. We accepted EST hits with the identity more than 95% and longer than 80% of EST sequence length as matches. For identification of PKs with overall high and low expression levels we selected genes with EST numbers >140 and <25, correspondingly, from pooled normal tissue EST libraries. For identification of PKs up-regulated and down-regulated in tissues, we normalized each EST library to the same level using published approach [63,65]. We calculated PK tissue expression score (ES) as the ratio of PK-specific ESTs versus expected EST frequency. We considered PKs with $ES > 6$ ($ES > 2$ for brain and

nervous tissue) as preferentially expressed in a tissue. PKs with $ES < 0.25$ from nervous tissue or testis were considered as down-regulated in these organs. Breadth of gene expression was estimated as the number of organ and tissue sources of gene-specific ESTs. For identification of the group of ubiquitously expressed PKs, we used tissue EST libraries containing more than 100,000 ESTs. Genes with low expression levels and low EST numbers, which could not be reliably evaluated by this method, were excluded from this analysis. Genes were ranked according to the number of expressing tissues. Genes expressed in 9 or more tissues from the 12 tissues were considered as broadly expressed. Groups of preferentially expressed PK genes were checked against the Gene Expression Atlas (<http://wombat.gnf.org/>) and available experimental PT-PCR and Northern data from literature. Only genes with similar tissue-specific preferences were considered in the final classification and computer analysis.

Textual and statistical analyses

Human-mouse evolutionary divergence of PK genes was evaluated using Kimura's two parameter model [28]. The levels of synonymous and non-synonymous divergence (K_s and K_a , respectively) were calculated with the PAML program (<ftp://abacus.gene.ucl.ac.uk/pub/paml>) using default parameters and the yn00 estimation method [25]. For all measures of evolutionary distances, including K_s , K_a , K_a/K_s , the Wilcoxon rank sum non-parametric test was applied to the pairwise comparison between all groups of PK genes.

To identify regulatory elements associated with transcript abundance and tissue-specific expression, we searched for conserved over-represented motifs in promoter regions of actively transcribed genes using the discriminating matrix emulator (DME) program [67]. Search for over-represented sequence elements in 5'UTR and 3'UTR regions was performed using an enumerative Markov chain motif finding algorithm [68], which applies z -scores to evaluate the over-representation of exact DNA words, and SiteDB program [69]. We also used the program CLOVER [70] that uses the position frequency matrices (PFMs) of cis-regulatory sites to evaluate sequences for statistically significant over/under-representative sequence elements. The methods employed take into account nucleotide content bias. Identified statistically significant over-represented motifs were compared with PFMs of known cis-regulatory motifs from the TRANSFAC database (<http://www.biobase-international.com/pages/index.php?id=transfac>) [71]. DiRE server for the identification of distant regulatory elements of co-regulated genes (<http://dire.dcode.org>) was used for prediction of transcription factor binding sites over-represented in conserved syntenic regions of PK genes predominantly expressed in nervous tissue [39].

Formation of intermolecular mRNA-rRNA duplexes and hybridization affinity of 5'UTRs to ribosomal RNA were evaluated with program Hybrid [62] under default parameters using ΔG threshold of ≤ -17 kcal/mol [57]. Annotated dataset of 476 human miRNAs was extracted from Rfam database, release 10 (<http://microrna.sanger.ac.uk/sequences/index.shtml>). For identification of potential miRNA target sites in 3'UTRs, we calculated hybridization affinity of miRNAs to 3'UTRs using Hybrid program and ΔG threshold of ≤ -17 kcal/mol, and used predictions of RegRNA program (<http://regrna.mbc.nctu.edu.tw/>). For identification of potential binding sites for neuron-specific miRNAs in 3'UTRs, we calculated hybridization affinity of 3'UTRs to annotated neuron-specific and brain-specific miRNAs from Rfam database. We identified common invariant oligonucleotides in 3'UTRs. We required common fragments of complementarity to be at least 6 nt long, since most identifies

targets have conserved complementary seeds of 6–8 nucleotides. We performed single-linking clustering of these targets using the Histogram AC program [69].

Statistical analysis was performed using exact Fisher test, Student's t-test for normally distributed variables, Wilcoxon rank sum test for unknown distributions.

Supporting Information

Table S1 PK-specific ESTs in GenBank originating from different human organs. Differentially expressed PK genes. Evolutionarily conserved motifs over-represented in promoter regions and 5'UTRs of high expression PK genes.

Found at: doi:10.1371/journal.pone.0003599.s001 (0.23 MB XLS)

Table S2 Top 25 transcription factor binding sites over-represented in evolutionarily conserved regions of PK genes preferentially expressed in the nervous tissue.

Found at: doi:10.1371/journal.pone.0003599.s002 (0.04 MB DOC)

Figure S1 Length of functional domains in the groups of differentially expressed human and mouse PK genes. A. Length of 5'-spacers, introns, and primary transcripts. B. Length of CDSs, 5'UTRs and 3'UTRs.

Found at: doi:10.1371/journal.pone.0003599.s003 (0.57 MB TIF)

References

- Hardie DG (1994) An emerging role for protein kinases: the response to nutritional and environmental stress. *Cell Signal* 6: 813–821.
- Hardie DG (2000) Metabolic control: a new solution to an old problem. *Curr Biol* 10: R757–759.
- Hunter T (2000) Signaling - 2000 and beyond. *Cell* 100: 113–127.
- Hanks S, Hunter T (1995) The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J* 9: 576–596.
- Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* 298: 1912–1934.
- Forrest AR, Ravasi T, Taylor D, Huber T, Hume DA, et al. (2003) Phosphoregulators: protein kinases and protein phosphatases of mouse. *Genome Res* 13: 1443–1454.
- Caenepeel S, Charyczak G, Sudarsanam S, Hunter T, Manning G (2004) The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc Natl Acad Sci USA* 101: 11707–11712.
- Quintaje SB, Orchard S (2008) The annotation of both human and mouse kinomes in UniProtKB/Swiss-Prot. *Mol Cell Proteomics* 7: 1409–1419.
- Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS (2001) Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet* 17: 373–376.
- Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520–562.
- Wasserman WW, Palumbo MWT, Fickett JW, Lawrence CE (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* 26: 225–228.
- Bernat JA, Crawford GE, Ogurtsov AY, Collins FS, Ginsburg D, et al. (2006) Distant conserved sequences flanking endothelial-specific promoters contain tissue-specific DNase-hypersensitive sites and over-represented motifs. *Hum Mol Genet* 15: 2098–2105.
- Kueng P, Nikolova Z, Djonov V, Hemphill A, Rohrbach V, et al. (1997) A novel family of serine/threonine kinases participating in spermiogenesis. *J Cell Biology* 139: 1851–1859.
- Tanaka H, Iguchi N, Nakamura Y, Kohroki J, de Carvalho CE, et al. (2001) Cloning and characterization of human haspin gene encoding haploid germ cell-specific nuclear protein kinase. *Mol Hum Reprod* 7: 211–218.
- Spiridonov NA, Wong L, Serfas PM, Starost MF, Pack SD, et al. (2005) Identification and characterization of SSTK: serine/threonine protein kinase essential for male fertility. *Mol Cell Biol* 25: 4250–4261.
- Wu JY, Ribar TJ, Cummings DE, Burton KA, McKnight GS, et al. (2000) Spermiogenesis and exchange of basic nuclear proteins are impaired in male germ cells lacking Camk4. *Nat Genet* 25: 448–452.
- Godbout M, Erlander MG, Hasel KW, Danielson PE, Wong KK, et al. (1994) IG5: a calmodulin-binding, vesicle-associated, protein kinase-like protein enriched in forebrain neurites. *J Neurosci* 14: 1–13.
- Dan C, Nath N, Liberto M, Minden A (2002) PAK5, a new brain-specific kinase, promotes neurite outgrowth in N1E-115 cells. *Mol Cell Biol* 22: 567–577.
- Theil T, Frain M, Gilardi-Hebenstreit P, Flenniken A, Charnay P, et al. (1998) Segmental expression of the EphA4 (Sek-1) receptor tyrosine kinase in the hindbrain is under direct transcriptional control of Krox-20. *Development* 125: 443–452.
- Takemoto-Kimura S, Terai H, Takamoto M, Ohmae S, Kikumura S, et al. (2003) Molecular cloning and characterization of CLICK-III/CaMKIgamma, a novel membrane-anchored neuronal Ca2+/calmodulin-dependent protein kinase (CaMK). *J Biol Chem* 278: 18597–18605.
- Sunyer T, Sahyoun N (1990) Sequence analysis and DNA-protein interactions within the 5' flanking region of the Ca2+/calmodulin-dependent protein kinase II alpha-subunit gene. *Proc Natl Acad Sci USA* 87: 278–282.
- Chen KH, Widen SG, Wilson SH, Huang KP (1993) Identification of a nuclear protein binding element within the rat brain protein kinase C gamma promoter that is related to the developmental control of this gene. *FEBS Lett* 325: 210–214.
- Dhariwala FA, Rajadhyaksha MS (2008) An unusual member of the Cdk family: Cdk5. *Cell Mol Neurobiol* 28: 351–369.
- Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA (2002) Selection for short introns in highly expressed genes. *Nat Genet* 31: 415–418.
- Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13: 555–556.
- Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* 17: 68–74.
- Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV (2004) Conservation and coevolution in the scale-free human gene coexpression network. *Mol Biol Evol* 21: 2058–2070.
- Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16: 111–120.
- Dresios J, Chappell SA, Zhou W, Mauro VP (2006) An mRNA-rRNA base pairing mechanism for translation initiation in eukaryotes. *Nature Struct Mol Biol* 13: 30–34.
- Shabalina SA, Ogurtsov AY, Spiridonov NA (2006) A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res* 34: 2428–2437.
- Shabalina SA, Spiridonov NA (2004) The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol* 5: 105.
- Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424: 147–151.
- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.

Figure S2 Correlation between PK expression levels and rates of non-synonymous human-mouse evolutionary divergence (Ka). Gene expression levels were estimated as the number of gene-specific ESTs in GenBank.

Found at: doi:10.1371/journal.pone.0003599.s004 (0.34 MB TIF)

Figure S3 Characteristic evolutionarily conserved motifs over-represented in promoter regions of high expression PK genes.

Found at: doi:10.1371/journal.pone.0003599.s005 (1.38 MB TIF)

Figure S4 Profiles of nucleotide base pairing in PK transcripts around the start codon (A) and the stop codon (B) with different mRNA structural domains. Blue, nucleotides paired with the 5'-UTRs; red, nucleotides paired with the CDSs; green, nucleotides paired with the 3'-UTRs; black, total base paired nucleotides.

Found at: doi:10.1371/journal.pone.0003599.s006 (0.88 MB TIF)

Acknowledgments

DME program used in this work was kindly provided by Andrew Smith. DiRE software was kindly provided by Ivan Ovcharenko. We thank anonymous reviewers for their helpful comments and suggestions.

Author Contributions

Conceived and designed the experiments: AYO SAS NAS. Performed the experiments: AYO LMR SAS. Analyzed the data: AYO LMR SAS NAS. Contributed reagents/materials/analysis tools: DL. Wrote the paper: GRJ SAS NAS.

34. Eisenberg E, Levanon EY (2003) Human housekeeping genes are compact. *Trends Genet* 19: 362–365.
35. Kochetov AV, Ischenko IV, Vorobiev DG, Kel AE, Babenko VN, et al. (1998) Eukaryotic mRNA encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS Lett* 440: 351–355.
36. Blanchette M, Bataille AR, Chen X, Poitras C, Laganier J, et al. (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* 16: 656–668.
37. Zhang C, Xuan Z, Otto S, Hover JR, McCorkle SR, et al. (2006) A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Res* 34: 2238–2246.
38. Kimmins S, Kotaja N, Davidson I, Saccone-Corsi P (2004) Testis-specific transcription mechanisms promoting male germ-cell differentiation. *Reproduction* 128: 5–12.
39. Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I (2007) Predicting tissue-specific enhancers in the human genome. *Genome Res* 17: 201–211.
40. Makalowski W, Boguski MS (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci USA* 95: 9407–9412.
41. Shabalina SA, Ogurtsov AY, Lipman DJ, Kondrashov AS (2003) Patterns in interspecies similarity correlate with nucleotide composition in mammalian 3'UTRs. *Nucleic Acids Res* 31: 5433–5439.
42. Mazumder B, Seshadri V, Fox PL (2003) Translational control by the 3'-UTR: the ends specify the means. *TiBS* 28: 91–98.
43. Mitchell P, Tollerway D (2001) mRNA turnover. *Curr Opin Cell Biol* 13: 320–325.
44. Balmer LA, Beveridge DJ, Jazayeri JA, Thomson AM, Walker CE, et al. (2001) Identification of a novel AU-rich element in the 3' untranslated region of epidermal growth factor receptor mRNA that is the target for repeated RNA-binding proteins. *Mol Cell Biol* 21: 2070–2084.
45. Shabalina SA, Koonin EV (2008) Origins and evolution of the eukaryotic microRNA systems. *Trends Ecol Evol* 23: 578–587.
46. Stark A, Brennecke J, Bushati NRBR, Cohen SM (2005) Animal microRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell* 123: 1133–1146.
47. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434: 338–345.
48. Nakayama T, Kuroi N, Sano MYT, Katsuya T, et al. (2006) Mutation of the follicle-stimulating hormone receptor gene 5'-untranslated region associated with female hypertension. *Hypertension* 48: 512–518.
49. Crocetto LE, Henderson BE, Coetzee GA (1997) Identification of two germline point mutations in the 5'UTR of the androgen receptor gene in men with prostate cancer. *J Urol* 158: 1599–1601.
50. Sgourou A, Routledge S, Antoniou M, Papachatzopoulou A, Psiouri L, et al. (2004) Thalassaemia mutations within the 5'UTR of the human beta-globin gene disrupt transcription. *Br J Haematol* 124: 828–835.
51. Borek G, Zarhrate M, Cluzeau C, Bal E, Bonnefont JP, et al. (2006) Father-to-daughter transmission of Cornelia de Lange syndrome caused by a mutation in the 5' untranslated region of the NIPBL Gene. *Hum Mutat* 27: 731–735.
52. Kozak M (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* 15: 8125–8148.
53. Kozak M (1991) An analysis of vertebrate mRNA sequences: intimations of translational control. *J Cell Biol* 115: 887–903.
54. Pickering BM, Willis AE (2005) The implications of structured 5' untranslated regions on translation and disease. *Seminars in Cell and Developmental Biology* 16: 39–47.
55. Signori E, Bagni C, Papa S, Primerano B, Rinaldi M, et al. (2001) A somatic mutation in the 5'UTR of BRCA1 gene in sporadic breast cancer causes down-modulation of translation efficiency. *Oncogene* 20: 4596–4600.
56. van der Velden AW, Thomas AAM (1999) The role of the 5' untranslated region of an mRNA in translation regulation during development. *Int J Biochem Cell Biol* 31: 87–106.
57. Matveeva OV, Shabalina SA (1993) Intermolecular mRNA-rRNA hybridization and the distribution of potential interaction regions in murine 18S rRNA. *Nucleic Acids Res* 21: 1007–1011.
58. Mauro VP, Edelman GM (2002) The ribosome filter hypothesis. *Proc Natl Acad Sci USA* 99: 12031–12036.
59. Chen LS, Tassone F, Sahota P, Hagerman PJ (2003) The (CGG)_n repeat element within the 5' untranslated region of the FMR1 message provides both positive and negative cis effects on in vivo translation of a downstream reporter. *Hum Mol Genet* 12: 3067–3074.
60. Ogurtsov AY, Roytberg MA, Shabalina SA, Kondrashov AS (2002) OWEN: aligning long collinear regions of genomes. *Bioinformatics* 18: 1703–1704.
61. Shabalina SA, Kondrashov AS (1999) Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet Res* 74: 23–30.
62. Nazipova NN, Shabalina SA, Ogurtsov AY, Kondrashov AS, Roytberg MA, et al. (1995) SAMSON: a software package for the biopolymer primary structure analysis. *Comput Appl Biosci* 11: 423–426.
63. Zhang Y, Eberhard DA, Frantz GD, Dowd P, Wu TD, et al. (2004) GEPIS: quantitative gene expression profiling in normal and cancer tissues. *Bioinformatics* 20: 2390–2398.
64. Zhang Y, Luoh SM, Hon LS, Baertsch R, Wood WI, et al. (2007) GeneHub-GEPIS: digital expression profiling for normal and cancer tissues based on an integrated gene database. *Nucleic Acids Res* 35: W152–158.
65. Zhu J, He F, Song S, Wang J, Yu J (2008) How many human genes can be defined as housekeeping with current expression data? *BMC Genomics* 9: 172.
66. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.
67. Smith AD, Sumazin P, Zhang MQ (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci USA* 102: 1560–1565.
68. Marino-Ramirez L, Spouge JL, Kanga GC, Landsman D (2004) Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res* 32: 949–958.
69. Kondrashov AS, Shabalina SA (2002) Classification of common conserved sequences in mammalian intergenic regions. *Hum Mol Genet* 11: 669–674.
70. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, et al. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* 32: 1372–1381.
71. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34: D108–110.

Research article

Open Access

Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites

Nak-Kyeong Kim, Kannan Tharakaraman, Leonardo Mariño-Ramírez and John L Spouge*

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Email: Nak-Kyeong Kim - kimnak@ncbi.nlm.nih.gov; Kannan Tharakaraman - tharakar@ncbi.nlm.nih.gov; Leonardo Mariño-Ramírez - marino@ncbi.nlm.nih.gov; John L Spouge* - spouge@ncbi.nlm.nih.gov

* Corresponding author

Published: 4 June 2008

Received: 29 October 2007

BMC Bioinformatics 2008, 9:262 doi:10.1186/1471-2105-9-262

Accepted: 4 June 2008

This article is available from: <http://www.biomedcentral.com/1471-2105/9/262>

© 2008 Kim et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Biologically active sequence motifs often have positional preferences with respect to a genomic landmark. For example, many known transcription factor binding sites (TFBSs) occur within an interval [-300, 0] bases upstream of a transcription start site (TSS). Although some programs for identifying sequence motifs exploit positional information, most of them model it only implicitly and with *ad hoc* methods, making them unsuitable for general motif searches.

Results: A-GLAM, a user-friendly computer program for identifying sequence motifs, now incorporates a Bayesian model systematically combining sequence and positional information. A-GLAM's predictions with and without positional information were compared on two human TFBS datasets, each containing sequences corresponding to the interval [-2000, 0] bases upstream of a known TSS. A rigorous statistical analysis showed that positional information significantly improved the prediction of sequence motifs, and an extensive cross-validation study showed that A-GLAM's model was robust against mild misspecification of its parameters. As expected, when sequences in the datasets were successively truncated to the intervals [-1000, 0], [-500, 0] and [-250, 0], positional information aided motif prediction less and less, but never hurt it significantly.

Conclusion: Although sequence truncation is a viable strategy when searching for biologically active motifs with a positional preference, a probabilistic model (used reasonably) generally provides a superior and more robust strategy, particularly when the sequence motifs' positional preferences are not well characterized.

Background

Transcription factor binding sites (TFBSs) provide a specific example of biologically functional sequence motifs that sometimes have positional preferences. TFBSs contribute substantially to the control of gene expression, and

because of their biological importance, much experimental effort has been expended in identifying them. Because experimental identification is expensive, there are now many computational tools that identify TFBSs as the sub-sequences, or "motifs", common to a set of sequences.

Most TFBSs correspond to short and imprecise motifs [1], however, so all computational tools in a recent contest performed rather poorly in identifying known TFBSs [2].

Although some tools have an *ad hoc* basis [3-5], other tools have a basis in the calculus of probability, and can therefore immediately and systematically combine sequence with other sources of information. Most probabilistic tools align candidate subsequences and convert the nucleotide counts in the alignment columns into a position-specific score matrix (PSSM). Most PSSMs are based on the log ratio between a motif model and a background model. Tools then identify putative motifs by maximizing the log ratio, usually with expectation maximization (EM) [6] or Gibbs sampling [7-9].

Experiments have shown, however, that besides common sequence motifs, TFBSs also have positional preferences, as illustrated in Figure 1. In yeast, TFBS positions demonstrate a strong bias toward locations between 150 and 50 bases upstream of the TSS [10]. In *E. coli*, TFBS positions tend to be located between 400 and 0 bases upstream of the translation start site [11]. In the words of Wray *et al.*, "for at least some regulatory elements, function constrains their position with respect to the transcriptional start site" (TSS) [1]. On the other hand, the trends regarding the positional preferences of TFBSs appear inconsistent. Wray *et al.* continue "for most transcription factors, however, binding sites lack any obvious spatial restriction relative to other feature of the locus" [1].

Some computational methods do exist to exploit the positional preferences of TFBSs. The first computational study using positional preferences used an empirical prior distribution of known positional information with respect to the translation start site from the *E. coli* genome [12]. This simple method, however, is applicable only to very simple

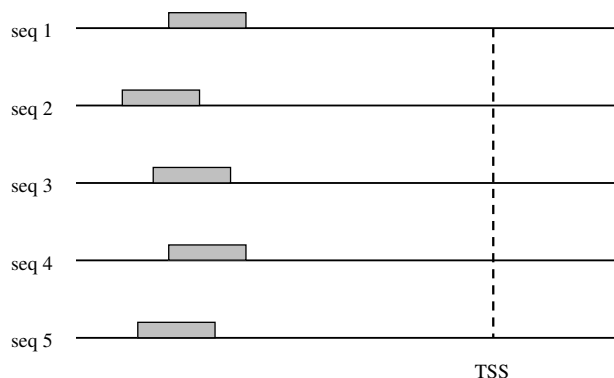


Figure 1
Positions of hypothetical TFBSs (gray boxes) with respect to the corresponding TSS.

organism like *E. coli*. Another computational study used position to calculate p-values for candidate motifs that formed a cluster [13]. The p-values were based on one particular database, however, and might not generalize reliably. Moreover, the corresponding model is not a probability model, making the systematic combination of sequence and positional information problematic. Yet another computational study modeled the positional preferences of TFBSs with a uniform prior, only mentioning the possibility of a more informative prior [11]. A systematic computational study to find new TFBS motifs by exploiting positional preferences applied a chi-square test to bins of positions near TSSs [14]. The chi-square test found one 8-letter word with significant positional preferences, the "Clus1" word, TCTCGCGA. The study's use of binning probably reduced the power of statistical tests, however. Shortly thereafter, in confirmation of the reduced statistical power, a systematic study of a human promoter dataset [15] identified 801 8-letter words with a positional preferences with respect to the TSS [9]. Interestingly, although 388 of the 801 words appeared in the TRANSFAC database [16], 413 of the words did not, suggesting that TFBS positional preferences were much more pervasive than previously believed. A later study showed that in eukaryotes the distribution of TFBSs was not uniform with respect to the TSS [17]. A study using chromatin immunoprecipitation followed by DNA hybridization (ChIP-Chip experiments) inferred TFBSs within sheared DNA fragments by using prior probability distributions to model positional preference [18]. The model was not directed at identifying TFBSs by their positional preferences with respect to genomic landmarks, however. Finally, a study applied a Poisson approximation to bins of positions within promoters to identify TFBSs by their positional preferences with respect to the TSS [19].

Several studies, therefore, have examined the positional conservation of TFBSs. Consequently, TFBS positional preferences are relatively well understood, particularly when compared to most non-coding DNA. Very few computational tools systematically combine positional preference with sequence information, however, and to our knowledge, no general-purpose computational tools using positional information are currently available. Standard tools like MEME [6], AlignACE [10], and Motif-Sampler [20], e.g., do not use positional information. Accordingly, this article evaluates the accuracy of predictions from a Bayesian model combining sequence with positional information, implemented in the newest version of the tool A-GLAM [9]. We assessed predictions from A-GLAM with and without the positional information, using a standard dataset of sequences with known TFBSs, and were therefore able to measure the contribution of positional information to TFBS prediction accuracy.

Results

Results for the TSS Tompa dataset

The TSS Tompa dataset is one of two test datasets considered in this study and contains 23 data subsets (see Methods). Table 1 shows an anecdotal A-GLAM alignment using positional information for the dataset 'hm08r' from the TSS Tompa dataset, which contains 10 sequences of length 2001. Run in its ZOOPS mode (Zero Or One Per Sequence), A-GLAM returned candidate alignments with only one or zero candidate site per sequence. In addition to sequence conservation, the alignment shows positional conservation within an interval of [-220, -1], much narrower than the input interval, [-2000, 0] bp upstream of the transcription start site (TSS). The alignment also overlapped several known sites (underlined in Table 1), with a correlation coefficient of 0.574, indicating good overlap.

Table 1 does not show the corresponding alignment without positional information, because its width was a biologically unrealistic 126 bp long. The alignment showed little positional conservation, with a range of [-2000, -1237]. It also showed essentially no overlap with the known sites, with a correlation coefficient of -0.012.

For TFBSs predicted without positional information, E-values were immoderately small, even for incorrect predictions. (Some incorrect predictions even displayed a numerical underflow E-value of 0, data not shown.) In contrast, the E-values in Table 1 were quite moderate, perhaps because they had to reconcile conflicting constraints from different sources of information on the motifs.

Alignments for more data subsets can be found in Supplementary Tables 1–6 [see Additional file 1]. We collected

Table 1: The A-GLAM output with positional information for 'hm08r'.

Name	Start	Alignment	End	Score	E-value
seq_0	-66	<u>GTCACGGC</u>	-59	11.0093	6.65E-06
seq_2	-65	<u>GTGACGTT</u>	-58	10.3315	2.30E-05
seq_3	-58	<u>ATGACGTC</u>	-51	11.2688	2.94E-06
seq_5	-188	<u>GTGACGTC</u>	-181	11.4594	1.28E-06
seq_7	-184	<u>CTGACGAC</u>	-177	9.86871	4.64E-05
seq_9	-101	<u>ATGACGTC</u>	-94	10.9283	8.09E-06
seq_10	-220	<u>ATCACGGC</u>	-213	7.58906	3.78E-04
seq_11	-80	<u>GTGACGTC</u>	-73	11.1306	4.75E-06
seq_12	-52	<u>CTGACGGC</u>	-45	10.0764	3.50E-05
seq_14	-8	<u>CTGATGTC</u>	-1	7.60515	3.69E-04

A-GLAM predicted TFBSs in 10 data subsets in the TSS Tompa data subset 'hm08r'. The column "Name" shows each data subset; the column "Alignment", the corresponding predicted TFBS. The start and end positions with respect to the corresponding TSS are shown in the columns "Start" and "End". The columns "Score" and "E-value" show bit scores and E-values that A-GLAM assigned to predicted TFBSs. The known binding sites in the alignment are underlined.

alignments (with positional information) whose correlation coefficient (CC) is larger than 0.08. The hm03r data subset does not appear in Tables 1–6, despite a CC of 0.386, because the corresponding alignment had a biologically unrealistic width of 224 bp. Unrealistically large alignment widths are much less common for alignments with positional information than without. In Supplementary Tables 2–6, the alignments without positional information are omitted because they show essentially no overlap with known binding sites.

Table 2 summarizes results for all 23 TSS Tompa data subsets. Some 18 out of the 23 datasets show improved predictions after adding positional information. Overall, the combined correlation coefficient (CCC; see Methods) at the bottom of Table 2 improved from -0.008 to 0.101. To evaluate the statistical significance, let γ and γ_+ denote the average correlation coefficient for each data subset without and with positional information. A one-sample Wilcoxon test against the one-sided null hypothesis $\gamma \geq \gamma_+$ yielded a p-value of 0.002, supporting the alternative hypothesis that $\gamma < \gamma_+$.

Results for TRANSFAC dataset

The TRANSFAC dataset contains 82 data subsets. Supplementary Table 8 contains detailed results for the input interval of [-2000, 0]. With the addition of positional information, the CCC has improved from -0.009 to 0.027 with a p-value of 10^{-8} (Wilcoxon test as above). The CCC for TRANSFAC dataset (0.027) is smaller than for TSS Tompa dataset (0.101), and the positional information makes a more significant change in the CCC for the TRANSFAC dataset ($p = 10^{-8}$) than for the TSS Tompa dataset ($p = 0.002$), probably because the TRANSFAC dataset contains 82 data subsets; the TSS Tompa dataset, only 23. In the case of subtle differences, the larger TRANSFAC dataset provides more evidence, leading to smaller p-values.

Cross-validation using TSS Tompa dataset

Because we used known binding sites to estimate the hyperparameters of the model (see Methods), one might suspect over-fitting. Moreover, because the distribution of locations might vary from one type of TFBS to another, the proposed model might not be appropriate for the discovery of unknown binding sites of different types of TFBSs. Cross-validation addressed these issues (see Methods).

Over the 100 random partitions from TSS Tompa dataset, the sample average of the CCC was 0.086; its sample standard deviation, 0.027; its 90% confidence interval, (0.049, 0.133); and its range, (0.029, 0.155). (The TRANSFAC dataset was not used for 5-fold cross-validation because of amount of computation required.) The

Table 2: The correlation coefficients for the TSS Tompa data subsets

Data Subset	Without positional information	With positional information	Improvement
hm01r	-0.012	-0.007	0.005
hm02r	-0.009	-0.007	0.002
hm03r	-0.037	0.386	0.423
hm04r	-0.008	-0.005	0.003
hm05r	-0.031	-0.019	0.012
hm06r	-0.014	0.156	0.170
hm07r	-0.015	-0.015	-0.001
hm08r	-0.012	0.574	0.586
hm09r	-0.011	0.358	0.369
hm10r	-0.019	0.083	0.102
hm11r	-0.028	-0.012	0.016
hm13r	-0.015	-0.016	-0.001
hm14r	0.204	-0.018	-0.222
hm15r	-0.011	-0.012	-0.002
hm16r	-0.011	-0.006	0.005
hm17r	-0.015	-0.012	0.004
hm18r	-0.018	0.094	0.112
hm19r	-0.010	-0.007	0.003
hm20r	-0.026	0.046	0.073
hm21r	0.401	0.384	-0.016
hm22r	-0.020	-0.020	0.000
hm24r	-0.016	-0.010	0.006
hm26r	-0.016	0.099	0.115
Combined CC	-0.008	0.101	0.109

Table 2 shows the correlation coefficients for A-GLAM's predictions on the 23 subsets of the TSS Tompa dataset. The column, "Improvement", quantifies the effect of positional information on predictions, by showing the difference between the correlation coefficients in the second and third columns, "Without Positional Information" and "With Positional Information".

CCC for the model using sequence information alone was -0.008. Because the CCC for sequence alone lay outside the range (0.029, 0.155) of the 100 CCCs using positional information in the 5-fold cross-validation, positional information improved prediction accuracy significantly. The actual CCC for the model using both sequence and positional information was 0.101 (see Table 2), well within the 90% confidence interval from cross-validation. The different types of known sites have quite diverse distributions (see Fig. 2), so we expect occasional misspecification of hyperparameters η in our model (see Methods). The 5-fold cross-validation shows, however, that classification accuracy is not excessively sensitive to the hyperparameter estimation or, by extension, to the locations of the known sites.

Truncation effect on sequences of test datasets

Figures 2 and 3 suggest that a truncated input sequence interval of, say, [-500, 0] or [-250, 0] might incorporate positional information as well as a Bayesian positional model applied to the full interval [-2000, 0]. Accordingly, in addition to the full interval [-2000, 0], we tested 3 truncated intervals [-1000, 0], [-500, 0], and [-250, 0]. (See Supplementary Table 7 and 8 for details.) The predictive accuracy, as represented by the CCCs in Table 3, indicate

that truncation on its own, without any Bayesian positional modeling, improved the motif predictions. Moreover, predictive improvements due to modeling position gradually disappeared as the truncation reduced the interval to [-250, 0]. Note, however, that positional modeling never significantly hurt the predictive accuracy, even with truncated input sequences.

Discussion

The new version of the A-GLAM program ('anchored gap-less local alignment of multiple sequences', written in C++) [9,21] can incorporate positional information by implementing the model from the Methods section in a Gibbs sampler. A-GLAM already has several desirable features when predicting transcription factor binding sites (TFBSs). First, it optimizes motif width automatically, without user input. Second, it reports theoretically accurate E-values for candidate TFBSs. Finally, it implements a theoretically sound context-dependent Markov background model, which yielded better predictions than different, *ad hoc* Markov background models or the conventional background model of independent bases [22]. With its Markov background model, a rigorous statistical evaluation showed that even before the addition of positional information, A-GLAM's predictive accuracy was

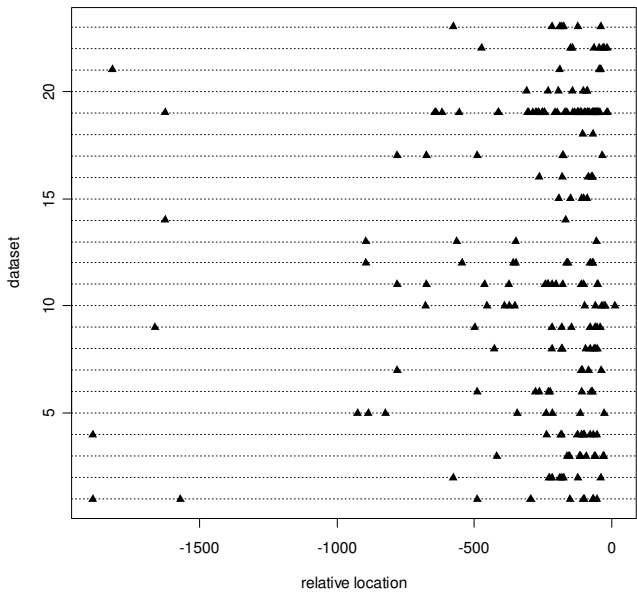


Figure 2
Distribution of known locations of binding site in TSS Tompa dataset. The x-axis is anchored on the TSS, denoted as location 0. All sequences in each test subset are collapsed into a single line; hence the 23 data subsets are shown as 23 different horizontal lines. Each data subset contains TFBSs corresponding to a single specific transcription factor.

competitive with any state-of-the-art motif-finding tool [22].

At the outset, we point out that all motif-finding tools have had notorious difficulty with the original Tompa dataset [2]. Our TSS Tompa test dataset is even more difficult than the original Tompa dataset. Its data subsets often contained fewer sequences than the corresponding origi-

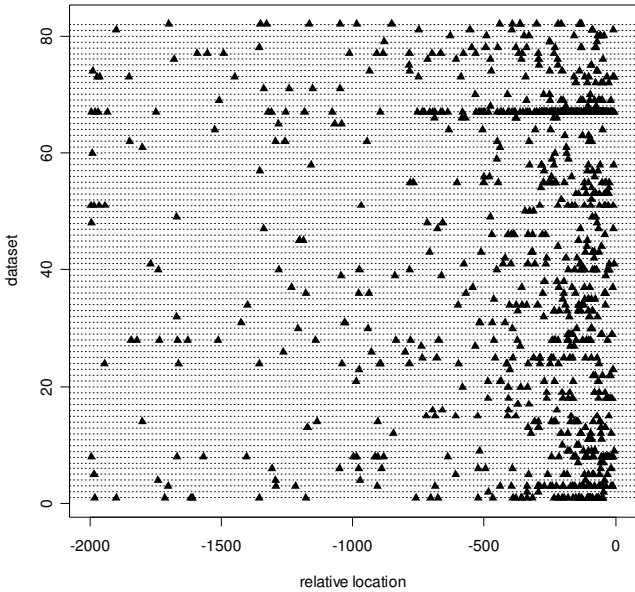


Figure 3
Distribution of known locations of binding site in TRANSFAC dataset. The x-axis is anchored on the TSS, denoted as location 0. All sequences in each test subset are collapsed into a single line; hence the 82 data subsets are shown as 82 different horizontal lines. Each data subset contains TFBSs corresponding to a single specific transcription factor.

nal Tompa subset. Moreover, our sequences were on average longer than the corresponding original Tompa sequence. Thus, conventional motif-finding tools should perform more weakly on our TSS Tompa test dataset than on the original Tompa dataset.

The Bayesian model in this paper combines sequence and positional information to predict putative TFBSs. Its implementation in A-GLAM permits users either to accept

Table 3: The effect of truncating the sequence upstream of the TSS

Sequence range	TSS Tompa Dataset			TRANSFAC Dataset		
	Without positional info	With positional info	p-value	Without positional info	With positional info	p-value
[-2000, 0]	-0.008	0.101	0.002	-0.009	0.027	10 ⁻⁸
[-1000, 0]	0.086	0.098	0.583	0.050	0.066	0.112
[-500, 0]	0.125	0.133	0.338	0.077	0.078	0.070
[-250, 0]	0.139	0.139	0.054	0.094	0.076	0.603

The first column shows the sequence range upstream of the TSS given as input to A-GLAM. The change of CCC from modes with and without positional information for the TSS Tompa and TRANSFAC datasets is displayed in the corresponding groups of three columns. The third column of each group shows a Wilcoxon p-value, which evaluates the difference between the CCCs in the previous two columns. Because not all TFBSs in our datasets are known, small improvements in the CCC correspond to true improvements of unknown magnitude. In particular, e.g., in the Table, two CCC values rounded to 0.139 have unseen decimals different enough to have a p-value of 0.054. To view results for individual sites in the Tompa dataset, see Supplementary Table 7 [see Additional file 1].

our default hyperparameters η for the prior distribution or to select their own. Although complete flexibility in the selection of hyperparameters can permit inappropriate or excessively aggressive choices, extensive cross-validation showed that the usual priors place mild restrictions on the predictions, so the model is very robust against misspecification of its hyperparameters or, by extension, to the locations of known sites. In other words, the prior does not dictate the alignment; instead, it loosely guides the alignment and permits the data to "speak for themselves". If motifs do not cluster by position, A-GLAM might therefore still find motifs sharing sequence but not position. We therefore make the following recommendation to users: in the absence of a strong reason to the contrary, they should accept A-GLAM's default hyperparameters.

To use positional information to find biologically active sites, A-GLAM's positional model requires the input sequences to be anchored on a genomic landmark, e.g., to find TFBSs, the model might be anchored to TSSs. Because a single gene might correspond to several alternative TSSs [23], however, TSS multiplicity might initially appear to cause problems. Moreover, the TSS itself can have either "sharp" or "broad" positional preference within a promoter [24]. Variability of the TSS position within a promoter reduces the positional information available to A-GLAM, possibly explaining the uneven improvement in prediction across our data subsets. A-GLAM's statistical model examines sequence as well as positional information, however, so it retains robustness against a mild misspecification of the TSS, say, within a few hundred bases of the true position, so alternative TSSs or TSSs with a typical broad positional distribution are unlikely to degrade predictions seriously when positional information is used. A-GLAM's users should note, however, if a TSS is specified, e.g., a kilobase away from the relevant position, positional information might severely distract A-GLAM from finding the desired TFBSs. On the other hand, however, different positions relative to the TSS containing exactly the same sequence have long been known to be associated with different TFBS biological functions [25]; in other cases, they might also be associated with alternative TSSs or TSSs with a broad positional distribution. Up to now, because computational studies of positional control of transcription have had to rely on *ad hoc* methods, A-GLAM now has a unique potential among general motif-prediction tools. Even if two functionally different sets of TFBSs have similar motifs, A-GLAM can differentiate them by position alone and report the two sets separately. It would be very interesting if someone using A-GLAM identified two sets of TFBSs of similar sequence corresponding to two different functionalities or TSSs.

The sequences in our study used the upstream positions from -2000 to 0 bp relative to the TSS to evaluate A-

GLAM's accuracy in predicting TSSs. Because our purpose in this article was to evaluate A-GLAM's ability to find biologically active sequence motifs in general, there is no scientific reason not to use the 3' UTR region as a "genomic anchor" to identify nearby regulatory elements. A similar statement applies to any set of regulatory elements (e.g., TFBSs, miRNA binding sites, etc.) around any genomic landmark (e.g., the TSS, the 3' UTR, etc.).

Indeed, if its main purpose was not evaluation of the predictive accuracy of A-GLAM's positional model, this article could have restricted its input sequences to intervals downstream of the TSS, e.g., [0, 1000] bp instead. With the TSS still providing the genomic anchor, A-GLAM could have searched for motifs associated with, e.g., 5' UTRs or translation start sites, which are usually within a few hundred base pairs downstream of a TSS. Thus, positional restrictions on the input sequence could focus A-GLAM's search on sequence motifs with different biological functions.

In practice, however, restricting the input interval requires great care. Unlike the TFBSs in our test datasets, many sequence motifs have poorly characterized distributions. On one hand, excessively stringent truncation of the input interval to, say, [-125, 0] would probably have removed many TFBSs from consideration in our study. On the other hand, positional modeling generally improved the accuracy of motif prediction, never hurting it significantly, even when input sequences were truncated. In the search for novel sequence motifs, therefore, we recommend that the use of Bayesian positional modeling on an input sequence whose length is generous (but not too generous) relative to the locations of known motifs.

Since the previous study showed that A-GLAM is one of the top performers among existing tools for *de novo* TFBS discovery [22], we believe that A-GLAM now easily outperforms its competitors whenever positional information is available and relevant. "Positional genomics" exploits the information provided by genomic landmarks (like the TSS), yielding a "poor man's alignment", even when the precise sequence alignments are unavailable. Given the power of comparative genomics, which depends on accurate alignments, positional genomics presents many interesting possibilities.

Conclusion

We proposed a Bayesian model for incorporating positional preference of TFBS with respect to a genomic landmark, e.g., a TSS. The results on our test datasets show that a positional model can produce statistically significant improvements in the accuracy of motif prediction. Our cross-validation study shows that the prior distribution of our positional model is robust against mild misspecifica-

tion of its parameters. Our study of truncated input sequences indicates that the positional model provides a superior and more robust strategy than sequence truncation, especially when the positional preferences of sequence motifs are not well characterized.

Availability

The A-GLAM program and all datasets relevant to this article can be found online [26].

Project name: A-GLAM 2.1

Project home page: <ftp://ftp.ncbi.nih.gov/pub/spouge/papers/archive/AGLAM/2008-02-20/>

Operating system: Linux

Programming language: C++

Licence: No license required.

Methods

The two test datasets

Our first test dataset was a subset of the "real" human sequences in the "original Tompa dataset", from [2]. The original Tompa dataset does not annotate any experimentally verified TSS positions, which were supplied from the Database of Transcription Start Sites (DBTSS) [27], as follows. BLAT [28] searched the DBTSS for hits to sequences in the original Tompa dataset. The DBTSS is incomplete, so when BLAT returned no hits in a sequence, the corresponding sequence was discarded. After the BLAT search, the dataset contained 26 data subsets, each composed of human sequences with a known TSS, and each corresponding to a single type of TFBS, like the original Tompa data subsets. We then discarded data subsets with 0 or 1 sequences, resulting in our "TSS Tompa dataset", which contained 23 data subsets. Each data subset contained from 2 to 26 sequences, and each sequence contained any number of known TFBSs, including 0. To encompass systematically all known TFBSs in the sequences, each sequence was expanded to contain proximal promoter regions from -2000 to 0 bp (upstream) relative to the corresponding TSS.

Our second test dataset was constructed from: (1) the latest human genome build (NCBI Build 36, ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/); (2) transcriptional start sites (TSS) from the database of transcription start sites (DBTSS) [27]; and (3) experimentally characterized TFBSs from the TRANSFAC database (professional version 11.2) [29]. Briefly, TSS and TRANSFAC sites were mapped to the human genome using MegaBLAST [30], yielding a set of proximal promoter DNA sequences [15,31] annotated with experimentally characterized TSSs

and TFBSs. In this paper, the resulting sequences are called our "TRANSFAC dataset". The TRANSFAC dataset contains 82 data subsets, each subset containing 2 to 101 sequences, and each sequence containing at least one instance of known TFBSs. Like the TSS Tompa data subsets, each data subset corresponded to a single type of TFBS. Like our TSS Tompa dataset, the range of TRANSFAC dataset is from -2000 to 0 bp (upstream) relative to the corresponding TSS.

A standard measure of prediction accuracy, the correlation coefficient, described elsewhere [22], evaluated TFBS predictions within our test dataset.

A Bayesian model for positional preferences

Our model for TFBSs uses two sources of information: sequence and position. We discuss sequence later, to focus on the novelties of position first.

Figure 2 displays the positions of all known TFBSs within the data subsets of the TSS Tompa dataset. Figure 2 collapses all sequences in each test subset into a single line anchored at the TSS. Thus, the 23 lines represent the 23 data subsets. Figure 2 shows that the TFBSs in several data subsets display positional preferences with respect to the TSS. Many TFBSs are upstream of the TSS, possibly clustered around certain positions. Accordingly, we search for TFBS positions that are normally distributed, with unknown center and dispersion, near the TSS. (Mathematical convenience facilitates the choice of the normal distribution.) Analogous to Figure 2, Figure 3 contains the positions of all known TFBSs in the TRANSFAC dataset. The TRANSFAC dataset displays the same basic distributional characteristics as the TSS Tompa dataset in Figure 2.

Fix a data subset in Figure 2 or 3, and assume it contains some number n of unknown TFBSs with locations x_1, \dots, x_n

relative to the TSS. For later reference, let $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$

and $s_n^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$ be the sample mean and sample standard deviation. Assume $\mathbf{x} = (x_1, \dots, x_n)$ constitute independent samples from a Normal (μ, λ) distribution, with mean μ and reciprocal variance (also known as "precision") $\lambda = 1/\sigma^2$. Given the normal parameters $\theta = (\mu, \lambda)$, the positions \mathbf{x} have the likelihood function

$$p(\mathbf{x} | \theta) = \left(\frac{\lambda}{2\pi} \right)^{n/2} \exp \left\{ -\frac{1}{2} \lambda \sum_{i=1}^n (x_i - \mu)^2 \right\}. \quad (1)$$

Parenthetically, to avoid confusion, the sequence locations x_1, \dots, x_n are integers, but the use of continuous distributions (e.g., the normal) as approximations simplifies

the algebra enormously. Similarly, the locations x_1, \dots, x_n might be confined to a finite interval (e.g., they might be within a finite piece of DNA). The seemingly unrestricted normal distribution remains appropriate, however, because its rapidly vanishing squared exponential form (as in Eq) effectively confines its samples to a finite interval.

Now, let the normal parameters $\theta = (\mu, \lambda)$ have a uniform-gamma prior distribution, in which μ and λ have independent prior distributions. The prior for μ is the continuous Uniform $[a, b]$ distribution on some closed interval $[a, b]$ ($a < b$), with constant density $p(\mu) = (b - a)^{-1}$ for $\mu \in [a, b]$. The prior for λ is a Gamma(α, β) distribution with parameters $\alpha, \beta > 0$, with density

$$p(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda)$$

for $\lambda \geq 0$. The uniform-gamma prior distribution for $\theta = (\mu, \lambda)$ therefore has the joint density function

$$p(\theta) = (b - a)^{-1} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda),$$

for $\mu \in [a, b]$ and $\lambda \geq 0$.

Practical suggestions for the numerical values of α and β are given below.

Our aim is to provide a figure of merit for Gibbs sampling based on the predictive distribution $p(x) = \int p(x|\theta)p(\theta)d\theta$ of the locations x . Gibbs sampling conditions on the locations $x = (x_1, \dots, x_n)$ to determine the conditional predictive distribution of the location $x_{n+1} = x$ of another TFBS (see Eq (2) below). After extensive algebraic manipulation of the relevant integrals, the conditional predictive distribution is

$$p(x|x) = \frac{p(x, x)}{p(x)} = \frac{\Gamma\left(\frac{1}{2}(v+1)\right)}{\Gamma\left(\frac{1}{2}v\right)} (v\pi)^{-1/2} \sigma^{-1} \left\{ 1 + \frac{(x - \bar{x}_n)^2}{v\sigma^2} \right\}^{-\frac{1}{2}(v+1)}, \quad (2)$$

a Student t-distribution whose parameters are $v = 2\left[\alpha + \frac{1}{2}(n-1)\right]$, \bar{x}_n , and

$$\sigma^2 + \frac{n+1}{n} \frac{\beta + \frac{1}{2}ns_n^2}{\alpha + \frac{1}{2}(n-1)}.$$

The t-distribution has mean \bar{x}_n for $v > 1$ and variance $v(v-2)^{-1}\sigma^2$ for $v > 2$.

The result in Eq (2) ignores the restriction $\mu \in [a, b]$. If $[a, b]$ covers most of the range $[a', b']$ of the locations x (e.g., $a-3\sigma < a' < b' < b+3\sigma$), then analysis will confirm that under appropriate mathematical hypotheses, Eq (2) approximates the desired conditional predictive distribution accurately.

The prior distribution is fully specified by a list of the hyperparameters a, b, α , and β . As indicated above, any sufficiently generous interval $[a, b]$ containing the locations x suffices for present purposes. The input sequence range (e.g., in the case of TSS Tompa's dataset as well as TRANSFAC dataset, from -2000 to 0 bp relative to the corresponding TSS) is a practical choice for $[a, b]$. In contrast, the selection of α and β can be delicate. On one hand, a user can provide subjective preferences for α and β , yielding a precision λ with mean $\alpha\beta^{-1}$ and variance $\alpha\beta^{-2}$. On the other hand, α and β can be estimated from the distributions of experimentally verified TFBSs, as follows.

Suppose we have k data subsets, where the i -th data subset ($i = 1, \dots, k$) yields a known vector x_i of locations for a particular TFBS. Each data subset x_i corresponds to a different set of hyperparameters $\{\theta_i = (\mu_i, \lambda_i)\}_{i=1, \dots, k}$ chosen from a common uniform-gamma prior with unknown parameters $\eta = (\alpha, \beta)$. The predictive distribution of the data is

$$p(x_1, \dots, x_k | \eta) = \left\{ \int p(x_1 | \theta_1) p(\theta_1 | \eta) d\theta_1 \right\} \dots \left\{ \int p(x_k | \theta_k) p(\theta_k | \eta) d\theta_k \right\}.$$

Maximization of the predictive distribution yields the so-called type-II maximum likelihood estimate for $\eta = (\alpha, \beta)$ [32].

In this study, based on our two datasets, the type-II maximum likelihood estimate of α and β were selected. The value of α was 0.8424; of β , 25790 for TSS Tompa dataset; α , 0.5825, β , 12818, for TRANSFAC dataset. Thus, the distribution of the precision λ had mean 3.27×10^{-5} (4.54×10^{-5} , for TRANSFAC dataset), giving the scale parameter $\sigma = \lambda^{-1/2}$ an approximate mean 175 (148, for TRANSFAC dataset). (The lengths of typical input sequences are several hundreds to a couple of thousand, e.g., in our dataset,

the lengths are all 2000.) Now, 95% of the realizations from a Normal (μ, λ) distribution fall into the interval $(\mu - 2\sigma, \mu + 2\sigma)$ of length 4σ . Because $4\sigma = 4(175) = 700$ (592, for TRANSFAC dataset), the above selection of α and β makes the prior distribution quite broad, permitting the data "to speak for themselves".

Some comments on the distributional choices for the prior and likelihood

The normal distribution might be challenged as an inappropriate form for the likelihood. In most of the data subsets in Figure 2 or 3, it is completely justifiable, but does appear untenable for a few. Although mathematical convenience facilitates the choice of a normal distribution, one could propose alternative distributional forms, usually at the expense of greater complexity. The normal distribution is quite adequate, however, when modeling any cluster lacking distant "orphan" locations.

Similarly, a uniform prior for the normal mean μ might be challenged. In fact, we implemented the same model with a normal-chi-square prior for $\theta = (\mu, \lambda)$. In our hands, both models produced comparable results on our test dataset (data not shown).

Gibbs sampling using both sequence and position

As noted above, Gibbs sampling requires only conditional predictive distributions. Because of the uniform prior for μ , multiplying the conditional predictive distribution in Eq (2) by (an ultimately irrelevant factor of) $(b - a)$ yields an approximation for the conditional predictive odds ratio with respect to the uniform background model. Taking logarithms and adding subscripts for "location", yields a log-odds score $\Delta_{s[l]}(x_{[l]} | x_{[l]})$ for location.

Now, consider the sequence information. Let the n locations $x_{[l]}$ initiate subsequences $x_{[s]}$ of length w (for "window"). Let the count of nucleotide j in the i -th column of the window be $c_{i,j}$, so the total count in each position is $c = \sum_{(j)} c_{i,j} = n$. As in the conditional predictive distribution above, add another subsequence $x_{[s]}$ of length w to the data. Let $\delta[i, j]$ equal 1 if the new subsequence contains nucleotide j in its i -th position, and 0 otherwise. Our previous work [9] postulated a familiar model [7,8], that the TFBS sequences follow a multinomial motif model with a Dirichlet prior. In the prior, the nucleotide pseudo-counts were $\{a_j\}$ ($a = \sum_{(j)} a_j$). The background model was the so-called "independent letters model" with probabilities $\{p_j\}$. Effectively, our previous work gave the conditional log-odds ratio of the subsequence $x_{[s]}$, given the subsequences $x_{[s]}$, as

$$\Delta_{s[s]}(x_{[s]} | x_{[s]}) = \sum_{i=1}^w \sum_{j=1}^4 \delta[i, j] \log \left[\left(\frac{c_{ij} + a_j}{c + a} \right) / p_j \right]. \quad (3)$$

If sequence and position are independent in both the motif and background models, the corresponding conditional predictive log-odds ratio is $\Delta_s(x | x) = \Delta_{s[s]}(x_{[s]} | x_{[s]}) + \Delta_{s[l]}(x_{[l]} | x_{[l]})$. Conditional predictive log-odds ratios can be added to generate the log-odds ratios for any dataset x step by step. Thus, Eqs (2) and (3) completely specify a predictive log-odds ratios for use as the figure of merit in Gibbs sampling. The present article actually replaces the independent letters model for the sequence background with a Markov model of order 3 [22], but the principles are the same.

Having established the separate roles of sequence and location, we drop the subscripts $[s]$ and $[l]$ below, particularly in x_i , which now represents the sequence and location of the i -th candidate TFBS.

A p-value for each candidate TFBS

For consistency with other computer programs (and because it makes little practical difference), to calculate a p-value for the i -th candidate TFBS x_i , we consider the self-predictive score $\Delta_s(x_i | x)$, where $x = (x_i, \dots, x_i, x_n)$ includes x_i . Because sequence and location are independent variates in both the motif and background models, the distribution of $\Delta_s(x_i | x)$ is a convolution, i.e.,

$$\mathbb{P}\{\Delta_s(x_i | x) \geq t\} = \sum_{(r)} \mathbb{P}\{\Delta_{s[s]}(x_{i,[s]} | x_{[s]}) = r\} \cdot \mathbb{P}\{\Delta_{s[l]}(x_{i,[l]} | x_{[l]}) \geq t - r\}$$

Existing methods [33,34] determine the distribution of $\Delta_{s[s]}$ and the distribution of $\Delta_{s[l]}$ is known. Thus, a p-value can be assigned to each candidate site.

k-fold cross-validation for sensitivity of hyperparameter selection

The k -fold cross-validation method estimates error rates in classification problems accurately [35]. The k -fold cross-validation splits the available data containing known classification labels into k mutually exclusive "partitions", so that each partition contains about the same amount of data. It then sets aside one of k partitions as the test set, and uses remaining $k - 1$ partitions as a training set to estimate the statistical parameters underlying the classification rule. After repeating the estimation process k times, leaving out each partition in turn, the average of the resulting classification errors estimates the error rate of the rule. The choice of 5 or 10 for k generally overcomes the

effects of replicated data, which would otherwise render the test and training data unduly similar [35]. In the present context, known sites provide estimates of the hyperparameters $\eta = (\alpha, \beta)$. In our study, cross-validation with $k = 5$ partitions was most appropriate to address over-fitting, because we have only 23 different datasets in TSS Tompa dataset. To illustrate the 5-fold cross-validation, consider the partition $23 = 5 + 5 + 5 + 4 + 4$. First, set aside the first "5" of the 23 data subsets as the test set x_1 , and estimate the hyperparameters η by maximizing the value of $p(x_2, \dots, x_5 | \eta)$, where x_2, \dots, x_5 are the $18 = 5 + 5 + 4 + 4$ training sets. With the estimated hyperparameters η , A-GLAM then makes predictions on the test set x_1 . The 5-fold cross-validation then repeats the procedure, taking each of the partitions x_2, \dots, x_5 in turn as the test set.

To eliminate the results' dependence on the partition, the partition was chosen randomly 100 times, and the results averaged.

A-GLAM Settings for the Test Predictions

To compare the model with positional information and the model without positional information (i.e., using sequence alone), we ran A-GLAM in the ZOOPS (Zero or One Occurrence Per Sequence) mode, where A-GLAM reports zero or one instance of the motif element for each sequence. Somewhat arbitrarily, we restricted the search space to the strands in the test dataset, without the complementary strands.

Authors' contributions

N-KK proposed the Bayesian model, implemented it in A-GLAM, and ran the program on the test datasets; KT and LM-R generated the test datasets and extracted the transcription start site information; JLS conceived and supervised the study.

Additional material

Additional file 1

Additional alignments for the TSS Tompa dataset and the complete data corresponding to the summary in Table 3. Supplementary Tables 1–6 contain additional alignments for the TSS Tompa dataset. Supplementary Table 7 summarizes truncation effects for the TSS Tompa dataset; Supplementary Table 8, for the TRANSFAC dataset.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-262-S1.doc>]

Acknowledgements

The authors thank Sergey Sheetlin for helpful discussion. This research was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health.

References

- Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20**(9):1377-1419.
- Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavese G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23**(1):137-144.
- Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15**(7-8):563-577.
- Pavese G, Mereghetti P, Mauri G, Pesole G: **Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes An algorithm for finding signals of unknown length in DNA sequences.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W199-203.
- Sinha S, Tompa M: **YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation.** *Nucleic Acids Res* 2003, **31**(13):3586-3588.
- Bailey TL, Elkan C: **Unsupervised learning of multiple motifs in biopolymers using expectation maximization.** *Machine Learning* 1995, **21**:51-83.
- Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262**(5131):208-214.
- Liu JS, Neuwald AF, Lawrence CE: **Bayesian models for multiple local sequence alignment and Gibbs sampling strategies.** *J Amer Statistical Assoc* 1995, **90**:1156-1169.
- Tharakaraman K, Marino-Ramirez L, Sheetlin S, Landsman D, Spouge JL: **Alignments anchored on genomic landmarks can aid in the identification of regulatory elements.** *Bioinformatics* 2005, **21**:1440-1448.
- Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*.** *J Mol Biol* 2000, **296**(5):1205-1214.
- Thompson W, Rouchka EC, Lawrence CE: **Gibbs Recursive Sampler: finding transcription factor binding sites.** *Nucleic Acids Res* 2003, **31**(13):3580-3585.
- McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.** *Nucleic Acids Res* 2001, **29**(3):774-782.
- Kielbasa SM, Korbel JO, Beule D, Schuchhardt J, Herzel H: **Combining frequency and positional information to predict transcription factor binding sites.** *Bioinformatics* 2001, **17**(11):1019-1026.
- FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C: **Clustering of DNA sequences in human promoters.** *Genome Res* 2004, **14**(15628):1562-1574.
- Marino-Ramirez L, Spouge JL, Kanga GC, Landsman D: **Statistical analysis of over-represented words in human promoter sequences.** *Nucleic Acids Research* 2004, **32**(3):949-958.
- Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31**(1):374-378.
- Li N, Tompa M: **Analysis of computational approaches for motif discovery.** *Algorithms Mol Biol* 2006, **1**:8.
- Qi Y, Rolfe A, MacIsaac KD, Gerber GK, Pokholok D, Zeitlinger J, Danford T, Dowell RD, Fraenkel E, Jaakkola TS, Young RA, Gifford DK: **High-resolution computational models of genome binding events.** *Nat Biotechnol* 2006, **24**(8):963-970.
- Defrance M, Touzet H: **Predicting transcription factor binding sites using local over-representation and comparative genomics.** *BMC Bioinformatics* 2006, **7**:396.
- Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y: **A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling.** *Bioinformatics* 2001, **17**(12):1113-1122.

21. Frith MC, Hansen U, Spouge JL, Weng Z: **Finding functional sequence elements by multiple local alignment.** *Nucleic Acids Res* 2004, **32(1)**:189-200.
22. Kim NK, Tharakaraman K, Spouge JL: **Adding sequence context to a Markov background model improves the identification of regulatory elements.** *Bioinformatics* 2006, **22(23)**:2870-2875.
23. Suzuki Y, Yamashita R, Nakai K, Sugano S: **DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs.** *Nucleic Acids Res* 2002, **30(1)**:328-331.
24. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38(6)**:626-635.
25. Ptashne M: **Lambda's switch: lessons from a module swap.** *Curr Biol* 2006, **16(12)**:R459-62.
26. John Spouge's Research Group [<http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/>]
27. Yamashita R, Suzuki Y, Wakaguri H, Tsuritani K, Nakai K, Sugano S: **DBTSS: DataBase of Human Transcription Start Sites, progress report 2006.** *Nucleic Acids Res* 2006, **34(Database issue)**:D86-9.
28. Kent WJ: **BLAT - The BLAST-like alignment tool.** *Genome Res* 2002, **12(4)**:656-664.
29. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34(Database issue)**:D108-10.
30. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** *J Comput Biol* 2000, **7(1-2)**:203-214.
31. Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK: **Transposable elements donate lineage-specific regulatory sequences to host genomes.** *Cytogenetic and genome research* 2005, **110(1-4)**:333-341.
32. Berger JO: **Statistical Decision Theory and Bayesian Analysis.** 2nd edition. New York, Springer-Verlag; 1985.
33. Huang H, Kao MC, Zhou X, Liu JS, Wong WH: **Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification.** *J Comput Biol* 2004, **11(1)**:1-14.
34. Kann MG, Sheetlin SL, Park Y, Bryant SH, Spouge JL: **The identification of complete domains within protein sequences using accurate E-values for semi-global alignment.** *Nucleic Acids Res* 2007, **35(14)**:4678-4685.
35. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning : data mining, inference, and prediction.** New York, Springer; 2001.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp



Research article

Open Access

Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA

Nalini Polavarapu¹, Leonardo Mariño-Ramírez², David Landsman², John F McDonald¹ and I King Jordan^{*1}

Address: ¹School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA and ²National Center for Biotechnology Information, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA

Email: Nalini Polavarapu - nalini@gatech.edu; Leonardo Mariño-Ramírez - marino@ncbi.nlm.nih.gov; David Landsman - landsman@ncbi.nlm.nih.gov; John F McDonald - john.mcdonald@biology.gatech.edu; I King Jordan^{*} - king.jordan@biology.gatech.edu

^{*} Corresponding author

Published: 17 May 2008

Received: 11 January 2008

BMC Genomics 2008, 9:226 doi:10.1186/1471-2164-9-226

Accepted: 17 May 2008

This article is available from: <http://www.biomedcentral.com/1471-2164/9/226>

© 2008 Polavarapu et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: The majority of human non-protein-coding DNA is made up of repetitive sequences, mainly transposable elements (TEs). It is becoming increasingly apparent that many of these repetitive DNA sequence elements encode gene regulatory functions. This fact has important evolutionary implications, since repetitive DNA is the most dynamic part of the genome. We set out to assess the evolutionary rate and pattern of experimentally characterized human transcription factor binding sites (TFBS) that are derived from repetitive versus non-repetitive DNA to test whether repeat-derived TFBS are in fact rapidly evolving. We also evaluated the position-specific patterns of variation among TFBS to look for signs of functional constraint on TFBS derived from repetitive and non-repetitive DNA.

Results: We found numerous experimentally characterized TFBS in the human genome, 7–10% of all mapped sites, which are derived from repetitive DNA sequences including simple sequence repeats (SSRs) and TEs. TE-derived TFBS sequences are far less conserved between species than TFBS derived from SSRs and non-repetitive DNA. Despite their rapid evolution, several lines of evidence indicate that TE-derived TFBS are functionally constrained. First of all, ancient TE families, such as MIR and L2, are enriched for TFBS relative to younger families like Alu and L1. Secondly, functionally important positions in TE-derived TFBS, specifically those residues thought to physically interact with their cognate protein binding factors (TF), are more evolutionarily conserved than adjacent TFBS positions. Finally, TE-derived TFBS show position-specific patterns of sequence variation that are highly distinct from random patterns and similar to the variation seen for non-repeat derived sequences of the same TFBS.

Conclusion: The abundance of experimentally characterized human TFBS that are derived from repetitive DNA speaks to the substantial regulatory effects that this class of sequence has on the human genome. The unique evolutionary properties of repeat-derived TFBS are perhaps even more intriguing. TE-derived TFBS in particular, while clearly functionally constrained, evolve extremely rapidly relative to non-repeat derived sites. Such rapidly evolving TFBS are likely to confer species-specific regulatory phenotypes, i.e. divergent expression patterns, on the human evolutionary lineage. This result has practical implications with respect to the widespread use of evolutionary conservation as a surrogate for functionally relevant non-coding DNA. Most TE-derived TFBS would be missed using the kinds of sequence conservation-based screens, such as phylogenetic footprinting, that are used to help characterize non-coding DNA. Thus, the very TFBS that are most likely to yield human-specific characteristics will be neglected by the comparative genomic techniques that are currently *de rigueur* for the identification of novel regulatory sites.

Background

The vast majority of the human genome is made up of non-protein-coding sequences [1,2], and the specific function of such DNA is often unknown. As of late, elucidating the functional relevance of the non-coding fraction of the human genome has become a major priority for computational and functional genomics [3].

Most of the non-protein-coding fraction of the human genome is made up of repetitive DNA sequences, primarily transposable elements (TEs), which alone make at least 45% of the genome. In one sense, these TEs can be considered as genomic parasites that exist solely by virtue of their ability to out-replicate the host genome in which they reside [4,5]. On the other hand, it has become abundantly clear that, once established in a genome, TEs can contribute to genome function in a number of different ways [6]. For instance, TEs are known to donate a wide variety of gene regulatory sequences to the human genome [7-9], and TE-derived regulatory sequences exert diversifying effects on the expression patterns of adjacent genes (reviewed in [10-12]).

TE-derived regulatory sequences are particularly interesting from an evolutionary perspective because of their potential to drive gene expression divergence between species. The potential for TEs to cause regulatory changes between evolutionary lineages is related to the fact that TEs invariably represent the most rapidly changing, lineage-specific part of eukaryotic genomes. For instance, when the human and mouse genomes sequences were compared, it became apparent that 99% of protein coding genes had human-mouse homologs, with 80% having direct 1:1 orthologs, whereas only 13% of mouse and 48% of human TEs were shared between the two species [13]. TE dynamics can even lead to substantial differences between genomes over relatively short evolutionary time scales. Indeed, the human evolutionary lineage has experienced a TE-driven genome expansion of 500 Mb in the last 50 million years and 30 Mb since the divergence from chimpanzees [14].

Taken together with their ability to donate regulatory sequences, this lineage-specific character of TEs suggests that the regulatory elements they donate may lead to species-specific differences in gene expression. In fact, a primate-specific endogenous retroviral element has been shown to donate an enhancer that confers a distinct parotid-specific expression pattern on the human amylase gene [15]. A more recent genome scale analysis showed that TE-derived human regulatory sites are associated with genes that have increased tissue-specific expression divergence between human and mouse [16]. A corollary prediction of this model for the diversifying regulatory effects of TEs is that TE-derived regulatory sequences will have

anomalously rapid evolutionary rates. Consistent with this expectation, we previously found that TE-derived human transcription factor binding sites (TFBS) are much less likely to have orthologs in the mouse genome than non-repetitive TFBS [17].

In this study, we set out to assess the relative evolutionary rates and the position-specific patterns of variation for human TFBS that are derived from repetitive versus non-repetitive DNA. We relied on the analysis of experimentally characterized TFBS that can be unambiguously mapped to the human genome in order to determine their evolutionary origins in repetitive or non-repetitive DNA. Our results suggest that TE-derived TFBS show both rapid evolution and, in some cases, anomalous position-specific patterns of change relative to non-repetitive TFBS. Despite these distinct evolutionary characteristics, the TE-derived TFBS do show sequence divergence patterns that are consistent with the conservation of function.

Results and Discussion

Human TFBS from repetitive DNA

A total of 2,521 experimentally characterized human TFBS were taken from the TRANSFAC database [18] and 1,810 of these were able to be precisely mapped to the latest build of the human genome reference sequence. Mapping of TFBS was done using the program site2genome, which facilitates unambiguous mapping of TFBS by using the longer flanking sequence context surrounding the relatively short binding sites [19]. The genomic locations of these human TFBS were compared to the locations of repetitive DNA sequences identified with the RepeatMasker program [20]. A total of 182 (10%) mapped human TFBS are co-located with repetitive DNA elements, and 121 (6.7%) of these are contained completely within repeats (Table 1). 62 of the TFBS derived completely from repeat regions are associated with TEs, while 59 are derived from simple sequence repeats (SSRs). SSRs are short tandem repeats consisting of repeated runs of exact or nearly exact k -mers, where $k = 1-13$ bp for microsatellites or $k = 14-500$ bp for minisatellites [1]. A lower percentage of the SSR co-located TFBS (57%) are found to completely overlap with the repeats compared to TE-derived TFBS (78%), suggesting that some of the SSR-derived TFBS identified here may represent ascertainment artifacts.

Human TEs can be characterized into specific classes/families, and the class/family-specific counts of TE-derived TFBS are shown in Table 1. The observed distributions of TE-derived TFBS across classes/families, relative to their expected distributions based on the genome frequencies of the TE classes/families, are shown in Figure 1. The human genome has experienced a number of successive waves of TE expansion, and accordingly, different TE fam-

Table 1: Counts for human TFBS derived from repetitive DNA.

Category	Total count	Complete overlap	Partial overlap
All repeats	182	121	61
All SSR	103	59	44
All TEs	79	62	17
Alu	20	19	1
MIR	16	10	6
L1	10	4	6
All other LINEs	10	8	2
LTR	14	14	0
DNA	9	7	2

The numbers of experimentally characterized TFBS mapped to different categories of human genome sequence are shown. Total counts are indicated along with counts for those cases where the TFBS completely or partially overlaps with the repeat.

ilies have distinct evolutionary ages [1]. For short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs), relatively older families, such as MIR and L2, encode more TFBS than expected based on their genome frequencies, while proportionally fewer TFBS are derived from younger element families such as Alu and L1. The relative enrichment of TFBS encoded by older TE families is consistent with the action of purifying selection based on their regulatory function. In other words, these older elements are likely to have been preserved in the genome because of the regulatory sequences that they provide as was predicted by Silva *et al.* [21].

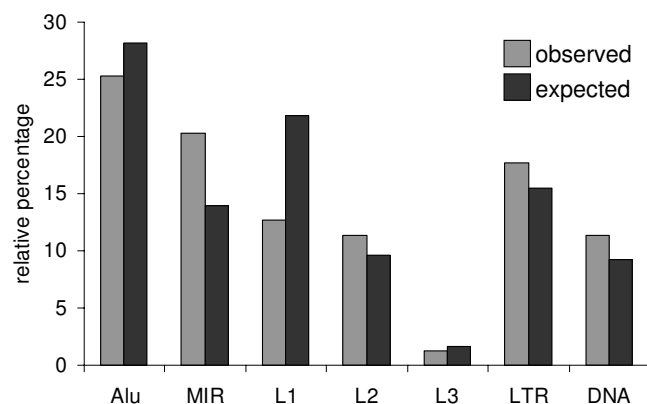


Figure 1
Observed versus expected frequencies of TE-derived TFBS. The observed percentages (light) of TE-derived TFBS from different classes/families of human TEs are plotted along with the percentages that are expected (dark) based on the background frequencies of the TEs in the genome. All class/family percentages are relative, i.e. they are normalized by the total number of TEs that donate TFBS (observed) and the total number of TEs in the genome (expected) respectively.

Evolutionary sequence conservation of repeat-derived TFBS

Levels of evolutionary sequence conservation between 17 vertebrate species were compared for TFBS with origins in repetitive versus non-repetitive DNA (Figure 2). TE-derived TFBS are by far the least conserved of the three categories, followed by SSR-derived and then non-repetitive TFBS. All differences between these categories are highly statistically significant ($110 > t > 19.0 = P < 9e-47$). This pattern of low sequence conservation for the TE-derived TFBS is consistent with the prediction of our regulatory divergence model that TEs are prone to provide rapidly evolving, lineage-specific TFBS.

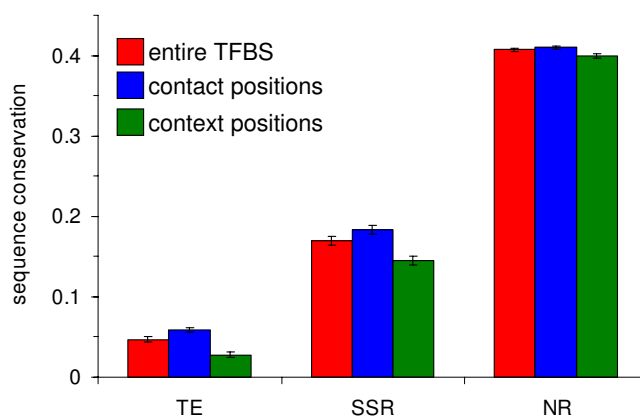


Figure 2
Average evolutionary sequence conservation for repetitive versus non-repetitive TFBS. Average conservation levels (\pm standard errors) are shown for TFBS that are derived from TEs, SSRs and non-repetitive DNA (NR). For each category, conservation levels were determined by averaging across the entire TFBS site (red), the specific contact part of the site that is thought to physically interact with the transcription factor (blue) and the sequence context part of the site that does contact the transcription factor (green).

Having shown the high levels of sequence divergence for TE-derived TFBS, it is worth noting that evolutionary conservation is often taken as a measure of functional relevance. For instance, the phylogenetic footprinting approach identifies highly conserved regulatory sequences as more likely to be functional [22,23]. While a number of functionally relevant TE-derived sequences have recently been identified by virtue of their sequence conservation [24-28], the relatively unconserved TE-derived TFBS revealed by our analysis would almost certainly be overlooked by phylogenetic footprinting methods. However, the TFBS that we analyzed were experimentally characterized, not predicted, and are thus quite likely to represent *bona fide* functional regulatory elements. In fact, the analysis of the relative evolutionary rates for different positions in the TFBS described below demonstrates that the specific pattern of conservation across sites supports the assertion that the TE-derived TFBS are functional.

TRANSFAC annotations in the site table represent individual residues in TFBS with either upper-case or lower-case letters. The upper-case residues correspond to specific sequence motifs within the site that were emphasized by the authors of the cited literature. We consider upper-case residues to be more likely to form specific DNA-protein contacts. Accordingly, the upper- and lower-case TRANSFAC annotations were used to partition TFBS residues into putative 'contact' positions, which are thought to physically interact with transcription factors (TF), versus 'context' positions that make up the rest of the site. Presumably, putative contact positions are more functionally relevant than context positions, *i.e.* a change of sequence at a contact position would have more of an effect on TF binding than a change at a context position would. If this is indeed the case, then according to the phylogenetic footprinting rationale, contact positions should be more conserved than context positions. This prediction is confirmed for all three categories of TFBS seen in Figure 2, and all differences between conservation levels for contact versus context positions within categories are statistically significant ($7.5 > t > 3.0$ $8.4 \times 10^{-11} < P < 2.5 \times 10^{-3}$). In other words, although TE-derived TFBS do evolve more rapidly than the other categories of TFBS, the position-specific patterns of TE-TFBS sequence divergence are nonetheless consistent with selective constraint based on their regulatory function.

Evolutionary conservation rates for contact and context positions were further broken down for the different classes/families of TEs (Table 2). These data reveal several noteworthy trends. There are substantial differences in the level of conservation among classes and families. For instance, it is not surprising that the evolutionarily young Alu family of elements has the least conserved TFBS, and

the young L1 family is similarly less conserved than the other older LINEs. One unexpected finding was the fact that TFBS derived from the long terminal repeats (LTRs) of endogenous retroviruses (ERVs) are the most conserved of all TE-derived TFBS. This observation stands out because ERVs are also evolutionarily young and not expected to be conserved. When this finding is considered together with the fact that LTRs are the only young class (or family) of TEs that has more TFBS than expected based on their genome frequencies (Figure 1), it suggests that LTRs may be particularly prone to donating regulatory sequences to the human genome. Indeed, LTRs are known to encode strong promoters, and there are a number of known cases where LTR-derived promoters control the expression of adjacent genes [29-33].

Another relevant point from the class/family specific evolutionary conservation data is the fact that the relative rates of contact versus context TFBS position divergence are consistent across all categories observed (Table 2). The greater conservation of contact positions is seen for even the least conserved Alu family ($t = 4.76$ $P = 2.7 \times 10^{-6}$). This indicates that the signal of functional constraint on TE-derived TFBS holds irrespective of the age of the elements from which the TFBS are derived, and serves as an independent confirmation of the experimental evidence in support of their identification.

Position-specific variation patterns for TE-derived TFBS

The results described in the previous section indicate that TE-derived TFBS show a low level of evolutionary conservation but a pattern of change that is consistent with their functional relevance as gene regulators. We used a probabilistic analysis of the position-specific patterns of sequence variation across TFBS sites to better understand the relative modes of evolution for non-repetitive versus

Table 2: Evolutionary sequence conservation of human TFBS.

Category	Site	Contact	Context
Non-repetitive	0.407 ± 0.085	0.410 ± 0.074	0.400 ± 0.110
All repeats	0.115 ± 0.042	0.130 ± 0.041	0.088 ± 0.045
All SSR	0.170 ± 0.056	0.183 ± 0.052	0.145 ± 0.062
All TEs	0.047 ± 0.026	0.059 ± 0.026	0.028 ± 0.026
Alu	0.002 ± 0.002	0.003 ± 0.003	0.002 ± 0.001
MIR	0.028 ± 0.017	0.048 ± 0.026	0.003 ± 0.004
L1	0.068 ± 0.063	0.077 ± 0.068	0.047 ± 0.052
All other LINEs	0.066 ± 0.018	0.095 ± 0.022	0.012 ± 0.011
LTR	0.141 ± 0.076	0.145 ± 0.042	0.136 ± 0.119
DNA	0.043 ± 0.029	0.057 ± 0.038	0.016 ± 0.009

Average (± standard deviation) base-by-base conservation levels are shown for different categories, non-repetitive and repetitive, of human TFBS. TFBS derived from repetitive DNA are broken down SSR versus TE-derived, and TE-derived TFBS are divided into specific classes/families of elements. Base-by-base conservation levels were averaged separately across entire sites, and across contact versus context positions.

TE-derived TFBS. To do this, position frequency matrices (PFMs) were taken from the TRANSFAC database for five TFBS where there was at least one TE-derived site in the human genome along with multiple non-repetitive TFBS. The PFMs summarize the collection of all experimentally characterized instances of that TFBS in the genome by representing the counts of each DNA residue (A, T, C or G) at each site in the TFBS (Figure 3). The PFMs can in turn be used to derive position weight matrices (PWMs), which are probabilistic representations of the position-specific nucleotide composition of the TFBS. The PWMs are represented as sequence logos [34], where the probabilities of observing a given residue at positions along the TFBS are indicated with the height of the residue symbols (Figure 3). We used these PWMs to score TE-derived versus non TE-derived TFBS sequences in terms of how well their specific sequences match the probabilistic model representing all other experimentally characterized sequences of that TFBS. The scoring was done using a 'leave-one-out' approach whereby each TFBS was scored using a PFM that does not include counts derived from the same TFBS. The TE-derived and non TE-derived sequence scores were compared to distributions of scores for three distinct simulated sets of 1,000 TFBS sequences. The first set of simulated TFBS sequences – 'genome-random' – was built by randomly drawing residues at each position of the TFBS based on their genome frequencies. The second set – 'repeat-random' – was generated from randomly sampled sequences, of the same length of the TFBS under consideration, taken from members of the same TE subfamily as the TE-derived TFBS being compared. Finally, the 'matrix-random' set was simulated according to the position-specific probabilities of the PWM for that TFBS.

An example of this kind of analysis can be seen for an Alu-derived TFBS (TRANSFAC site R08639) that sits just upstream of the FOS-like antigen (FOSL1)-encoding gene on human chromosome 11 (Figure 4). This TFBS was identified by virtue of its interaction with the beta-catenin-T cell-factor/lymphoid-enhancer-factor complex (Tcf/Lef) [35]. In that same study [35], binding of Tcf/Lef to FOSL1 and C-JUN was implicated in the progression of colon carcinoma. Interestingly, both FOSL1 and C-JUN are part of the AP-1 transcription complex suggesting that this Alu-derived TFBS may be involved in a cascade of regulatory interactions.

The particular TRANSFAC PFM model that corresponds to this Alu-derived site is M00671, and the binding factor for this model is the T-cell-specific transcription factor 4 (TCF-4 aka TCF7L2). The PFM and derived PWM that correspond to the M00671 model are shown in Figure 3. This PWM was used to calculate scores for sets of genome-random, repeat-random and matrix-random sequences (Figure 5A). The Alu-derived and the non-repetitive TFBS were

scored using PWMs built from M00671 PFMs that do not include residue counts from the particular TFBS being scored, *i.e.* using the leave-one-out method (Figure 5B). As could be expected, the genome-random and repeat-random simulated TFBS sequences have lower scores than do the matrix-random simulated sequences (Mann-Whitney U test $P = 3.7 \times 10^{-5}$). What is more relevant is the fact that all of the experimentally characterized TFBS have scores that fall within the range of the matrix-simulated sequences and are much higher than either the genome-random or repeat-random scores (Table 3). This includes the Alu-derived TFBS, which scores significantly higher than the average scores for the genome-random and repeat-random sites (Mann-Whitney U test $P = 1.9 \times 10^{-3}$). In other words, the Alu-derived TFBS has a position-specific DNA sequence profile that much more closely resembles the non TE-derived sites than it resembles random genomic sequences or random Alu sequences of the same subfamily. However, the Alu-derived site does have a lower score than all of the other non TE-derived sites. This indicates that there is still something unique about the TE-derived site relative to the non TE-derived sites. Thus, the position-specific profile of the Alu-derived TCF-4 binding site shows the hallmark of being functionally active yet retains a unique character relative to the non TE-derived sites that bind the same factor. The four other sites analyzed here show similar patterns in that they are clearly non-random, *i.e.* they score higher than the genome-random and repeat-random sets, and thus appear to be functional (Table 3). For the p53 matrix (M00761) Androgen receptor matrix (M00962), the TE-derived sites score lower than the non-repetitive sites; the two other cases show TE-derived sites with higher average scores than the non-repetitive sites. However, these differences are not statistically significant, indicating that TE-derived TFBS have position-specific profiles that are indistinguishable from non-repetitive TFBS. This is consistent with the fact that we started with experimentally characterized TFBS and underscores the functional relevance, and similar position-specific evolutionary constraints, of these TE-derived TFBS.

Conclusion

There are numerous experimentally characterized TFBS in the human genome (7–10%) that are derived from repetitive DNA indicating a pronounced effect of repetitive DNA on human gene regulation. TFBS that originate from repeats evolve more rapidly than non-repetitive TFBS but still shown signs of sequence conservation on functionally critical residues due to purifying selection. Position-specific patterns sequence variation observed for TE-derived TFBS, in terms of the specific nucleotide composition along the positions of the TFBS, also point to divergence in the face of functional constraint. These findings are consistent with the notion that TFBS originating from

M00671 TCF-4

	A	C	G	T
01	1	3	2	0
02	0	6	0	0
03	0	1	0	5
04	0	0	0	6
05	0	0	0	6
06	0	0	5	1
07	6	0	0	0
08	3	0	1	2



M00761 TP53

	A	C	G	T
01	25	3	16	2
02	14	0	32	0
03	25	0	21	0
04	2	39	4	1
05	32	2	4	8
06	23	2	2	19
07	3	0	43	0
08	9	15	5	17
09	2	28	9	7
10	5	22	5	14



M00789 GATA

	A	C	G	T
01	50	8	8	39
02	1	0	103	1
03	104	0	1	0
04	0	0	0	105
05	89	1	3	12
06	58	3	39	5
07	28	18	48	11



M00962 AR

	A	C	G	T
01	11	2	4	13
02	1	5	23	1
03	26	0	1	3
04	8	1	20	1
05	4	22	1	3
06	29	1	0	0
07	8	12	5	5
08	11	0	13	6
09	9	6	6	9



M01037 GLI1

P0	A	C	G	T
01	4	5	1	5
02	0	5	5	5
03	0	1	0	14
04	0	0	12	3
05	0	0	15	0
06	0	0	15	0
07	2	0	1	12
08	0	0	14	1
09	0	0	13	2
10	1	1	1	12
11	0	12	2	1
12	1	7	4	3



Figure 3
Probabilistic modelling of TFBS. PFM for five collections of human TFBS (Table 3) are shown along with sequence logo representations of their PWMs. Each PFM/PWM represents a human TFBS that has both TE-derived and non-repetitive experimentally characterized sites in the genome. The TFBS are identified with their TRANSFAC matrix identifiers and the official human gene name symbol for the binding transcription factor proteins.

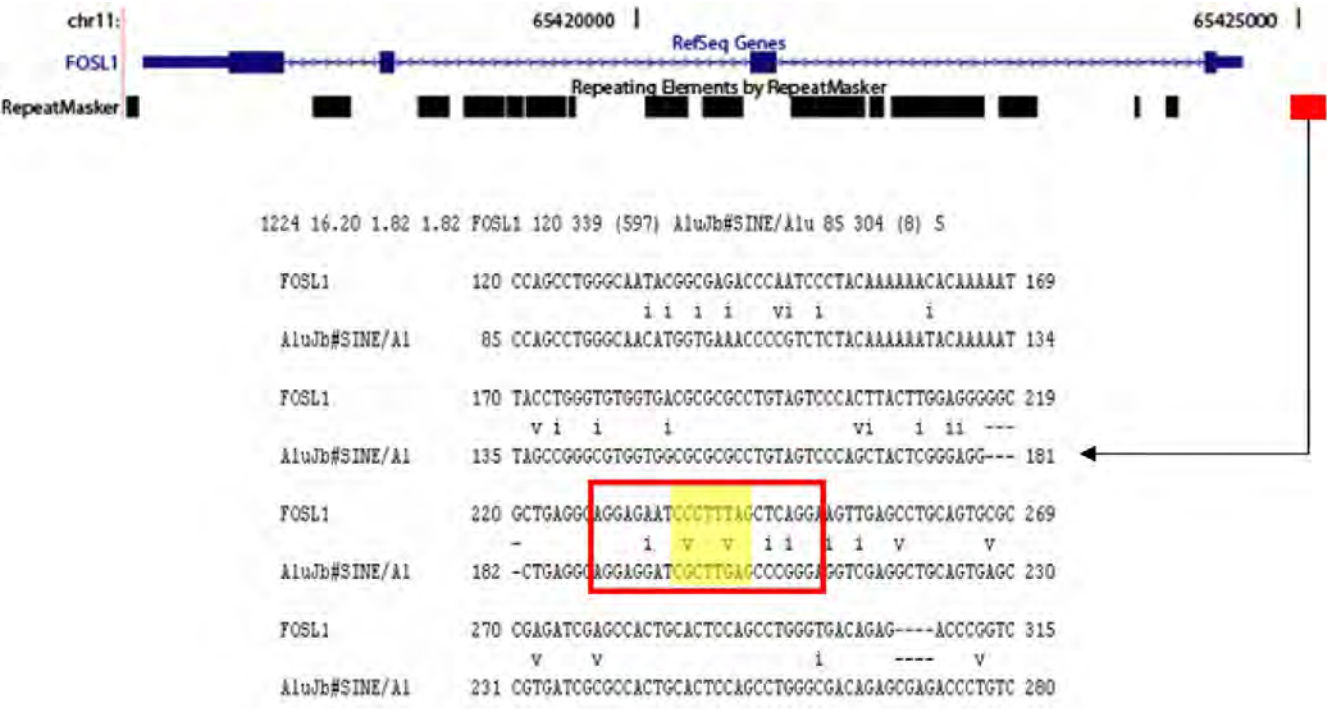


Figure 4
An Alu-derived TFBS upstream of the FOSL1 encoding gene. A schematic of the intron-exon structure of FOSL1, taken from the UCSC genome browser, is shown (blue) along with the positions of the repetitive DNA elements (black) at that locus. FOSL1 is encoded on the Crick strand of human chromosome 11. An Alu insertion (red) that donates a TCF-4 binding sites is found just upstream of the FOSL1 5' untranslated region in the proximal promoter region. Summary statistics and a sequence alignment between the FOSL1 proximal promoter sequence and the AluJb subfamily consensus sequence are shown with the TFBS location indicated (entire site boxed in red, contact residues highlighted in yellow).

repetitive DNA elements are likely to provide functionally relevant regulatory divergence between species.

transposable elements (TEs) and simple sequence repeats (SSRs), annotated with the RepeatMasker program [20].

Methods

Experimentally characterized human transcription factor binding sites (TFBS) were retrieved from the Professional release 11.3 (9/10/07) of the TRANSFAC database [18]. These TFBS were mapped to the July 2003 human reference sequence [1] (National Center for Biotechnology (NCBI) Build 34 or hg16) using the program site2genome [19]. For many individual TFBS, TRANSFAC annotations list GenBank accessions that provide longer flanking sequence context for the relatively short TFBS contained within the sequence. Site2genome uses this flanking sequence context to allow for one-to-one TFBS-to-genome mapping. Only TFBS that could be unambiguously mapped to the human genome sequence (1,810 out of 2,521) were taken for further analysis, and these TFBS mappings were transferred to the current human genome build (NCBI Build 36 or hg18) using the UCSC Genome Browser [36] 'lifter' utility. The locations of human TFBS were compared to the locations of repetitive DNA,

The evolutionary conservation levels for human TFBS were determined based on complete genome sequence alignments [37] between the human genome and 16 other vertebrate genomes [38]. These alignments have been analyzed, along with the phylogenetic tree of the species, by the program phastCons [39] to make predictions of discrete conserved genomic elements and to produce conservation level scores for each position (base) in the human genome. The base-by-base conservation level scores range from 0 to 1 and represent the posterior probability of every individual position in the genome being in a conserved element. Base-by-base conservation level scores were taken across all positions of the mapped TFBS and then averaged for the different categories compared in Table 2 and Figure 2.

Individual TFBS were broken down into putative contact and context positions using the TRANSFAC site table annotations. In the site table, the TFBS sequences are rep-

Table 3: Position-specific sequence variation scores for TE-derived, non-repetitive, matrix-random and genome-random TFBS.

Matrix ¹	Protein binding factor ²	TE-derived	non-repetitive	matrix-rand	genome-rand	repeat-rand
M00671	T-cell-specific transcription factor 4 (TCF-4 or TCF7L2)	4.25	5.69 ± 0.51	5.80 ± 0.73	-48.76 ± 15.61	-48.63 ± 14.97
M00761	p53 (TP53)	5.97	6.65 ± 1.26	5.52 ± 1.92	-2.79 ± 3.02	-4.71 ± 3.35
M00789	GATA binding proteins (GATA)	6.12	5.26 ± 1.56	5.27 ± 1.46	-5.87 ± 2.71	-4.70 ± 3.15
M00962	Androgen receptor (AR)	3.72	4.45 ± 1.21	4.33 ± 1.74	-2.29 ± 1.28	-1.80 ± 2.17
M01037	Glioma-associated oncogene homolog 1 (GLI1)	9.34	9.12 ± 1.14	9.24 ± 1.70	1.77 ± 2.83	-4.28 ± 2.91

Average TFBS scores are shown for each category of sites.
¹The TRANSFAC database matrix identifier
²The colloquial name of the protein that binds the TFBS along with its official HUGO name in parentheses

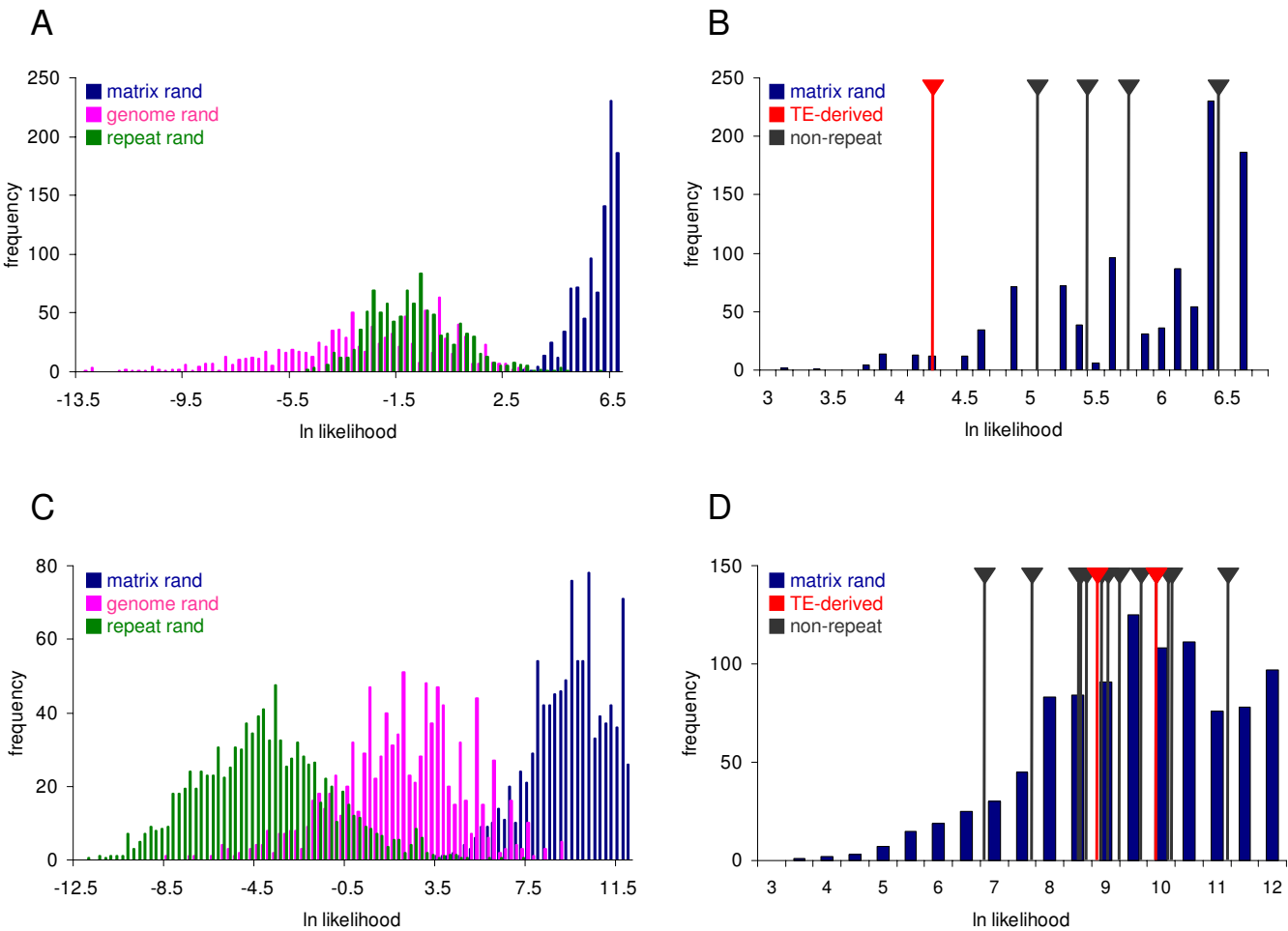


Figure 5
Site-specific variation scores for TE-derived versus non-repetitive TFBS. (A & C) Frequency distributions of scores for 1,000 simulated genome-random sequences (pink), repeat-random sequences (green) and matrix-random sequences (blue) for the M00671 matrix representing the TFBS bound by TCF-4 (A) and the M01037 matrix for TFBS bound by GLI1 (C). (B & D) The matrix-random score distributions are compared to the scores for individual TFBS derived from TEs (red) versus the non-repetitive TEs (gray). Data are shown for M00671 TCF-4 (B) and M01037 GLI1 (D).

resented with upper-case and lower-case residues. The upper-case TFBS residues correspond to specific sequence motifs within the site that were emphasized by the authors of the cited literature. We consider upper-case residues to be more likely to form specific DNA-protein contacts than lower case residues. Accordingly, the upper- and lower-case TRANSFAC annotations were used to partition TFBS residues into putative 'contact' positions, which are thought to physically interact with transcription factors (TF), versus 'context' positions. TFBS were also divided into those derived from repetitive, TE and SSR, versus non-repetitive classes and average conservation scores were determined for each TFBS class over each residue (contact and context) class. The statistical significance of the differences between average evolutionary conservation levels was evaluated using the Students' t-test.

Analysis of the site-specific pattern of TFBS evolution was done using probabilistic models of TFBS that were computed based on a previously described protocol [40]. Position frequency matrices (PFMs), which represent the counts of each of the four DNA residues (A, T, C and G) in each position of a TFBS model, were downloaded from TRANSFAC 10.3. PFMs were converted into position-weight matrices (PWMs), which represent the probability (p) of observing each DNA residue (r) at each position (i) in a TFBS according to the following formula:

$$p_{r,i} = \frac{c_{r,i} + s_r}{n + 4 * s_r}$$

where $c_{r,i}$ = counts of residue r at position i , s_r is a pseudo-count function = 1, and n = the total number of TFBS used to build the model. These probabilities ($p_{r,i}$) are normalized by the background genome frequencies of the DNA residues (p_r) to compute weights (W):

$$W_{r,i} = p_{r,i} / p_r$$

The PWMs are represented as sequence logos [34], which were built from the collections of TFBS sequences provided by the TRANSFAC matrix database, using the program WebLogo [41]. PWMs were used in Monte-Carlo simulation to build test sets of 1,000 TFBS sequences, the so-called 'matrix-random' sequences. For this procedure, DNA residues at each position of a TFBS were drawn at random according the site-specific probabilities of its PWM. 'Genome-random' simulated sets of 1,000 TFBS were built by randomly drawing residues across site positions according to their background genome frequencies. 'Repeat-random' simulated sets of 1,000 TFBS were generated by randomly sampling sequences of the same length of the matrix from members of the same repeat (TE) sub-family that the particular TE-derived TFBS was derived.

The PWMs were used compute scores (S) individual observed and simulated TFBS according to the formula:

$$S = \sum_{i=1}^n \ln W_{r,i}$$

where $W_{r,i}$ = the weight of the observed residue r at position i and n = the number of sites in the TFBS PWM. Individual TFBS from the TRANSFAC site table were scored using the leave-one-out method whereby matrix-specific PFMs were iteratively built without residue counts from the particular TFBS being scored. Scores (S) were compared for individual TE-derived and non-repetitive TFBS along with the score distributions for simulated sets of matrix-random and genome-random sites.

Authors' contributions

IKJ and LM-R conceived of and designed the study and performed computational analyses. LM-R and DL provided data used for the computational analyses. NP performed computational analyses in the lab of JFMCD. IKJ drafted the manuscript. All authors read and approved of the manuscript.

Acknowledgements

IKJ was supported by the School of Biology at the Georgia Institute of Technology. LM-R and DL were supported by the Intramural Research Program of the National Center for Biotechnology Information, National Library of Medicine at the National Institutes of Health. JFMCD and NP were supported by a grant from the Georgia Tech Research Foundation.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409(6822)**:860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al.: **The sequence of the human genome.** *Science* 2001, **291(5507)**:1304-1351.
3. Consortium EP: **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306(5696)**:636-640.
4. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution.** *Nature* 1980, **284(5757)**:601-603.
5. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite.** *Nature* 1980, **284(5757)**:604-607.
6. Kidwell MG, Lisch DR: **Perspective: transposable elements, parasitic DNA, and genome evolution.** *Evolution Int J Org Evolution* 2001, **55(1)**:1-24.
7. Jordan IK, Rogozin IB, Glazko GV, Koonin EV: **Origin of a substantial fraction of human regulatory sequences from transposable elements.** *Trends Genet* 2003, **19(2)**:68-72.
8. Thornburg BG, Gotea V, Makalowski W: **Transposable elements as a significant source of transcription regulating signals.** *Gene* 2006, **365**:104-110.
9. Lagemaat LN van de, Landry JR, Mager DL, Medstrand P: **Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions.** *Trends Genet* 2003, **19(10)**:530-536.
10. Britten RJ: **DNA sequence insertion and evolutionary variation in gene regulation.** *Proc Natl Acad Sci USA* 1996, **93(18)**:9374-9377.
11. Britten RJ: **Mobile elements inserted in the distant past have taken on important functions.** *Gene* 1997, **205(1-2)**:177-182.
12. Medstrand P, Lagemaat LN van de, Dunn CA, Landry JR, Svenback D, Mager DL: **Impact of transposable elements on the evolution**

- of mammalian gene regulation. *Cytogenet Genome Res* 2005, **110**(1-4):342-352.
13. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al.: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**(6915):520-562.
 14. Liu G, Zhao S, Bailey JA, Sahinalp SC, Alkan C, Tuzun E, Green ED, Eichler EE: **Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome.** *Genome research* 2003, **13**(3):358-368.
 15. Samuelson LC, Wiebauer K, Snow CM, Meisler MH: **Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution.** *Mol Cell Biol* 1990, **10**(6):2513-2520.
 16. Marino-Ramirez L, Jordan IK: **Transposable element derived DNaseI-hypersensitive sites in the human genome.** *Biol Direct* 2006, **1**:20.
 17. Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK: **Transposable elements donate lineage-specific regulatory sequences to host genomes.** *Cytogenet Genome Res* 2005, **110**(1-4):333-341.
 18. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al.: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31**(1):374-378.
 19. Frith MC, Halees AS, Hansen U, Weng Z: **Site2genome: locating short DNA sequences in whole genomes.** *Bioinformatics* 2004, **20**(9):1468-1469.
 20. RepeatMasker [<http://www.repeatmasker.org/>]
 21. Silva JC, Shabalina SA, Harris DG, Spouge JL, Kondrashovi AS: **Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes.** *Genet Res* 2003, **82**(1):1-18.
 22. Gumucio DL, Heilstedt-Williamson H, Gray TA, Tarle SA, Shelton DA, Tagle DA, Slightom JL, Goodman M, Collins FS: **Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes.** *Mol Cell Biol* 1992, **12**(11):4919-4929.
 23. Zhang Z, Gerstein M: **Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements.** *J Biol* 2003, **2**(2):11.
 24. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D: **A distal enhancer and an ultraconserved exon are derived from a novel retroposon.** *Nature* 2006, **441**(7089):87-90.
 25. Kamal M, Xie X, Lander ES: **A large family of ancient repeat elements in the human genome is under strong selection.** *Proc Natl Acad Sci USA* 2006, **103**(8):2740-2745.
 26. Lowe CB, Bejerano G, Haussler D: **Thousands of human mobile element fragments undergo strong purifying selection near developmental genes.** *Proc Natl Acad Sci USA* 2007, **104**(19):8005-8010.
 27. Nishihara H, Smit AF, Okada N: **Functional noncoding sequences derived from SINEs in the mammalian genome.** *Genome research* 2006, **16**(7):864-874.
 28. Xie X, Kamal M, Lander ES: **A family of conserved noncoding elements derived from an ancient transposable element.** *Proc Natl Acad Sci USA* 2006, **103**(31):11659-11664.
 29. Bannert N, Kurth R: **Retroelements and the human genome: new perspectives on an old relation.** *Proc Natl Acad Sci USA* 2004, **101**(Suppl 2):14572-14579.
 30. Dunn CA, Medstrand P, Mager DL: **An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon.** *Proc Natl Acad Sci USA* 2003, **100**(22):12841-12846.
 31. Dunn CA, Romanish MT, Gutierrez LE, Lagemaat LN van de, Mager DL: **Transcription of two human genes from a bidirectional endogenous retrovirus promoter.** *Gene* 2006, **366**(2):335-342.
 32. Romanish MT, Lock WM, Lagemaat LN van de, Dunn CA, Mager DL: **Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution.** *PLoS Genet* 2007, **3**(1):e10.
 33. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D: **Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53.** *Proc Natl Acad Sci USA* 2007, **104**(47):18613-18618.
 34. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Res* 1990, **18**(20):6097-6100.
 35. Mann B, Gelos M, Siedow A, Hanski ML, Gratchev A, Ilyas M, Bodmer WF, Moyer MP, Riecken EO, Buhr HJ, et al.: **Target genes of beta-catenin-T cell-factor/lymphoid-enhancer-factor signaling in human colorectal carcinomas.** *Proc Natl Acad Sci USA* 1999, **96**(4):1603-1608.
 36. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome research* 2002, **12**(6):996-1006.
 37. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al.: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome research* 2004, **14**(4):708-715.
 38. **Vertebrate Multiz Alignment & Conservation (17 Species)** [<http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=100603286&c=chrX&g=multiz17way>]
 39. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al.: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome research* 2005, **15**(8):1034-1050.
 40. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat Rev Genet* 2004, **5**(4):276-287.
 41. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome research* 2004, **14**(6):1188-1190.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp



The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site

Kannan Tharakaraman¹, Olivier Bodenreider², David Landsman¹,
John L. Spouge¹ and Leonardo Mariño-Ramírez^{1,*}

¹Computational Biology Branch, National Center for Biotechnology Information and ²National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, MSC 6075 Bethesda, MD 20894-6075, USA

Received February 14, 2008; Revised March 11, 2008; Accepted March 12, 2008

ABSTRACT

A number of previous studies have predicted transcription factor binding sites (TFBSs) by exploiting the position of genomic landmarks like the transcriptional start site (TSS). The studies' methods are generally too computationally intensive for genome-scale investigation, so the full potential of 'positional regulomics' to discover TFBSs and determine their function remains unknown. Because databases often annotate the genomic landmarks in DNA sequences, the methodical exploitation of positional regulomics has become increasingly urgent. Accordingly, we examined a set of 7914 human putative promoter regions (PPRs) with a known TSS. Our methods identified 1226 eight-letter DNA words with significant positional preferences with respect to the TSS, of which only 608 of the 1226 words matched known TFBSs. Many groups of genes whose PPRs contained a common word displayed similar expression profiles and related biological functions, however. Most interestingly, our results included 78 words, each of which clustered significantly in two or three different positions relative to the TSS. Often, the gene groups corresponding to different positional clusters of the same word corresponded to diverse functions, e.g. activation or repression in different tissues. Thus, different clusters of the same word likely reflect the phenomenon of 'positional regulation', i.e. a word's regulatory function can vary with its position relative to a genomic landmark, a conclusion inaccessible to methods based purely on sequence. Further integrative analysis of words co-occurring in PPRs also yielded 24 different groups of genes, likely identifying

cis-regulatory modules *de novo*. Whereas comparative genomics requires precise sequence alignments, positional regulomics exploits genomic landmarks to provide a 'poor man's alignment'. By exploiting the phenomenon of positional regulation, it uses position to differentiate the biological functions of subsets of TFBSs sharing a common sequence motif.

INTRODUCTION

In the postgenomic era, the identification of signals regulating transcription remains an outstanding problem (1,2). The problem has frustrated standard methods in computational sequence analysis, and experiments still provide one of the few consistently reliable sources of information about transcriptional signals (3). Even simple *cis*-regulatory transcription-binding sites (TFBSs) have proved notoriously difficult to identify *de novo*, because they usually correspond to short, degenerate motifs whose sequence information is insufficient on its own for dependable predictions. In particular, sequence analysis alone is generally unable to address the information that higher-order chromatin structure contributes to gene regulation (4).

Consider, however, a transcriptional complex anchored on a transcription start site (TSS). Each transcription factor (TF) within the complex occupies a particular position. Thus, if a TF interacts with a TFBS, the TFBS probably is constrained positionally with respect to the TSS. Moreover, as classic experiments on the lambda repressor and its operator-binding sites showed, by occupying TFBSs in different positions, a single TF can assume different biological functions (5). Rather like a receptor antagonist occupying a binding site, a TFBS

*To whom correspondence should be addressed. Tel: +301 402 3708; Fax: +301 480 2288; Email: marino@ncbi.nlm.nih.gov

corresponding to the TF might activate in one position relative to the TSS, but repress in another. Because of position, therefore, a single TFBS motif might regulate gene expression in a tissue- or temporal stage-specific manner (or both). Positional regulation of function generalizes obviously and broadly, to regulatory elements and genomic landmarks other than TFBSs and TSSs.

In the presence of positional regulation, sequence alone would be insufficient to predict TFBS function. Fortunately, many modern databases annotate their sequences. Consequently, where the traditional conference slide in computational biology once displayed an endless sea of letters, it should now display letters punctuated regularly by genomic landmarks like the TSS, exon boundaries, etc. Presently, genomic investigations are not exploiting the position of annotated landmarks as much as they might.

Positional regulomics therefore holds promise, but it requires in hand a rich source of interesting regulatory positions. With regard to TFBSs, some computational studies have examined position (6–10), but few new putative motifs emerged. In contrast, our previous work discovered 791 eight-letter DNA words displaying positional preferences with respect to the TSS (11). To summarize the work, the Database of Transcription Start Sites (DBTSS) contained many human TSSs determined from oligo-capping experiments (12–14). False positive TSSs were eliminated by precise transcript mapping, yielding a database of 4737 putative promoter regions (PPRs) containing positions –2000 to +1000 bp relative to the corresponding TSS (15). For each of the 4^8 eight-letter DNA words, a local maximum statistic (similar to the BLAST statistic) assessed the word's positional preferences with respect to the TSS (11). After multiplying by 4^8 to correct for multiple testing, the analysis yielded 791 statistically significant words ($P \leq 0.05$). Of the 791 words, 388 had perfect matches in TRANSFAC database (16), an event with a P -value of 4×10^{-42} . The biological function of the other 413 of the 791 words remained unidentified, but suggested the potential of positional regulomics to discover unknown sequence elements and their function.

To give an overview of the present study, with recent TSS data (17), the PPR dataset now contains 7914 sequences (see the Methods section). Within the new PPR dataset, the local maximum statistic identified words w displaying positional preferences with respect to the TSS. (To avoid unnecessary repetition of the phrase 'with respect to the TSS', all bp coordinates and positions below refer implicitly to the corresponding TSS, unless stated otherwise.) Each statistically significant positional preference yielded a 'cluster' of positions containing the corresponding word w , and each of the clusters corresponded to a group of genes ('gene group'). Occasionally, a single word w corresponded to more than one cluster, hinting at the possibility of a TFBS under positional regulation, and rendering such words particularly interesting to us.

Two external sources of information implicated the positional clusters in the co-regulation of the corresponding gene group. First, quantitative functional relationships were determined using a semantic similarity method (18)

based on the Gene Ontology (GO) annotation. The functional analysis suggested that many individual gene groups had a common biological function. Second, the microarray experiments in the GNF Atlas 2 (19) suggested that many individual gene groups identified here were co-expressed across multiple tissues. In addition to validating the biological functionality of words and helping to classify the corresponding putative TFBS, the two sources of information permitted us to formulate some novel biological hypotheses. In accord with the notion of positional regulation, our analysis sometimes linked different tissues to *specific positions* of a word, to our knowledge yielding the first computational evidence that a TFBS's position can influence the tissue-specificity of its regulatory functions. Furthermore, in accord with the analogy to receptor antagonists, our analysis sometimes linked different levels of activation or repression of the same gene group in different tissues to specific positions of a word. Thus, it is not an isolated phenomenon in human gene regulation, that the position of a TFBS influences its function in a regulatory module.

METHODS

The PPR database

Recently, (17) determined new TSSs with about 1.8 million 5'-end clones of full-length human cDNAs, extending the DBTSS. DBTSS yielded 30 924 TSSs for 14 628 RefSeq (20) human genes, indicating that many genes have alternative TSSs. The PPR database was constructed using every TSS within ± 1000 bp of the start of an annotated RefSeq transcript for annotated genes. If several TSSs were within ± 1000 bp of the same RefSeq 5' end, the closest TSS was used. The corresponding PPRs in DBTSS were aligned to the human genome (NCBI, build 36). Each PPR that mapped unambiguously was extended to include from –2000 to +1000 bp relative to the TSS (which was at 0 bp), as in our previous study (11). The final PPR database contained 7914 sequences. An ungapped block alignment then anchored the PPRs, placing all TSSs in a single column. Supplementary Figure S1 shows systematic variations in base composition over the alignment columns, confirming that the anchored alignment generally placed the TSSs correctly.

Our previous study describes in detail the remaining procedures, applied to every one of the 4^8 eight-letter DNA words. For each word and for each PPR, one instance of the word was chosen uniformly at random, and the remaining instances masked. At the end of the masking procedure, each PPR contained at most one unmasked instance of the word, in a random position. The unmasked instances in each column of the block alignment were counted, and a local maximum statistic (similar to the gapless BLAST statistic) assessed whether the unmasked instances of the word were unusually clustered by columns within the block alignment (see Supplementary data—Section 1.1). The randomized masking step reduces the density of ubiquitous repetitive elements or low complexity regions (e.g. poly A, poly T), which are biologically uninteresting in the present context but which

tend to be statistically significant without masking. Our study examined clusters with a significant local maximum statistic, a ‘cluster’ being simply a statistically significant set of positions within certain PPRs.

Pairwise correlation coefficient for microarray data

Given the set of $n = 74$ tissue-specific microarray expression values (X_i, Y_i) for two genes g_1 and g_2 , the corresponding Pearson correlation coefficient is

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}. \quad 1$$

Pairwise correlation coefficient for significant words

Each significant word W provided a pairwise similarity corresponding to TFs in TRANSFAC (16), as follows.

We used 522 count matrices from TRANSFAC Professional 11.1, many of which represent the same or similar factors. To make the set nonredundant, we skipped all nonvertebrate matrices, and if a family of related factors shared a single matrix, the matrix appeared once, to represent the entire family. For each of the 145 nonredundant count matrices remaining, the standard log-likelihood ratio yielded a PSSM as follows: Let p_n represent the background probability of nucleotide $n \in \{a, c, g, t\}$ in the 7914 PPRs. Let $c_{n,k}$ represent the count of nucleotide n in column k . Then, the score for nucleotide i at column k is

$$s_{n,k} = \ln \left[\left(\frac{c_{n,k} + a_n}{c + a} \right) / p_n \right], \quad 2$$

where $c = \sum_{n \in \{a, c, g, t\}} c_{n,k}$ is the total number of counts, which is independent of the column k (being the total number of TFBSs in TRANSFAC corresponding to the TF in question); a_n is the pseudo-count, which regularizes count matrices based only on a few TFBSs; and $\alpha = \sum_{n \in \{a, c, g, t\}} \alpha_n$. As in previous studies (11), we took $a_n = 1.5 \times p_n$.

With the 145 nonredundant PSSMs in hand, we calculated match scores for each word W and PSSM M , as follows: Each PSSM was padded with eight columns of 0s on each side. As above, let $s_{n,k}$ denote the score for of nucleotide n in column k , where $k = -8, \dots, -1, 0, \dots, j-1, j, \dots, j+7$, the columns $k = 0, \dots, j-1$ being from the original PSSM. The word $W = W(0), \dots, W(7)$ receives a maximum score

$$S_{M,W} = \max_{i=0, \dots, j+7} \sum_{a=0}^7 s_{W(a), i-8+a}. \quad 3$$

The summed score on the right of Equation (3) can be related to the binding energy of the TF for the putative TFBSs (21). The maximum score $S_{M,W}$ is the best summed score that the word W receives in any offset against PSSM M .

With the maximum scores $S_{M,W}$ in hand, we calculated empirical P -values for each word W from our significant

clusters, as follows. For each PSSM M , all eight-letter words yielded 65 536 maximum scores $S_{M,W}$ against the PSSM. For any word W , consider the corresponding maximum score $S_{M,W}$. The empirical P -value $p_{M,W}$ for the sequence W against the PSSM M is the fraction of the 8-mers that have a maximum score higher than $S_{M,W}$. The complement $1 - P_{M,W}$ of the P -value then should increase with the binding energy for the word W and the TF generating the PSSM M . The complement $1 - P_{M,W}$ is also normalized between 0 and 1.

Now, let $i = 1, \dots, 145$ index the nonredundant PSSMs M_i ; and let $g = 1, \dots, 3589$ index the genes in our dataset. If the words W_1, \dots, W_w correspond to the gene with index g , define $T_{g,i} = \max_{w=1, \dots, w} (1 - P_{i,w})$ ($i = 1, \dots, 145$) if $w > 0$ and 0 otherwise.

In the table $\{T_{g,i}\}$, the rows represent the genes; the columns, TFs. When Equation (1) is applied to $X_i = T_{g_1, i}$ and $Y_i = T_{g_2, i}$, which correspond to the genes g_1 and g_2 , it yields the Pearson correlation coefficients (PCCs) between rows in the table $\{T_{g,i}\}$. Two genes therefore receive a high PCC, if they correspond to similar words, regardless of the words' positions. The resulting network still reflects putative TFBSs as predicted by positional preference, however.

The integration of positional, functional and co-expression data

Three networks were constructed using positional, functional and co-expression data. In the corresponding networks, an edge joined a gene pair, if the pair scored above the 95th percentile for the corresponding measure: (i) 0.566 for the Pearson correlation coefficient quantifying TF positional similarity; (ii) 0.588 for the GO semantic similarity or (iii) 0.546 for the PCC from the microarray Atlas data. The networks were analyzed using Cytoscape (22), software freely available from <http://www.cytoscape.org> for visualizing molecular interaction networks. The Graph Merge plug-in, also freely available from <http://www.cytoscape.org>, produced the intersection network (see Supplementary Figure S3) whose edges lie in all three networks. Supplementary Table 2 lists numbers of nodes and edges, and average degrees for each of the four networks. The MCODE Cytoscape plug-in (23) identified 24 densely connected sets of genes in the intersection network.

RESULTS

Many words displaying positional preferences are probably functional

After multiplying P -values by 4^8 to correct for multiple testing, our methods yielded 1226 eight-letter words with significant positional preferences ($P \leq 0.05$). Out of the 1226 words, 71 words corresponded to two significant clusters with distinct positions and seven words corresponded to three significant clusters with distinct positions, for a total of 1311 significant clusters. (To avoid unnecessary repetition, all the ‘words’, ‘clusters’, and ‘gene groups’ mentioned below are significant at $P \leq 0.05$ after multiple test corrections, unless stated otherwise.)

Supplementary data file 1 contains the words and their clusters. To identify similar or overlapping words, we varied one base within each word, but few words were similar or overlapped. Only 608 of the 1226 words exactly matched subsequences of experimentally determined TFBSs in the TRANSFAC database. To discover relationships between the words and basal promoter elements like the CAAT box, SP1, CREB and TATA box (recognized by the constitutive human factors NF-Y, SP1, CREB and TBP, respectively), we again varied one base within each word. With a change of at most one base, 540 of the 1226 words exactly matched a consensus subsequence of one of the basal promoter elements. Because the regions surrounding many human genes are GC-rich, we examined the sequence composition of the words. Within the 1226 words, the frequencies of A, C, G and T were 0.159, 0.318, 0.376 and 0.146, respectively. Moreover, 123 words ($\approx 10\%$) contained only G and C, but only eight words ($\approx 0.65\%$) contained only A and T. Thus, the words do indeed reflect the elevated GC content around the TSS (Supplementary Figure S1).

Our previous study only found TFBSs from -200 bp to $+100$ bp relative to the TSS at 0 bp. Moreover, to permit a genome-scale study, our methods here ruthlessly sacrificed statistical power in favor of computational speed, so they probably found a small fraction of all functional TFBSs (which our unpublished data estimates loosely at a site-level sensitivity of about 15%). As expected, clusters upstream of the TSS were all within -200 bp relative to the TSS, indicating that our methods do not find TFBSs distant from the TSS. Clusters downstream of the TSS usually occurred within $+100$ bp. Cluster density peaked roughly at the TSS. Some 44 clusters were positioned more than $+100$ bp downstream of the TSS. A consensus GT dinucleotide appeared in 22 of the corresponding words, suggesting their role in mRNA splicing.

Many clusters correspond to gene groups with a common function

We investigated the (significant) gene groups for common functions, analyzing annotations from the GO database (24). Although several tools for analyzing GO annotations are publicly available (25–27), none was entirely suitable for our study, so we developed other methods ourselves.

Accordingly, we used semantic similarity measures to quantify the commonalities of molecular function for each pair of the 15 536 *Homo sapiens* gene products with GO annotations (18) (see Supplementary data—Section 1.3). The semantic measures yielded a maximum average pairwise functional similarity (APFS) within each gene group. Similarly, we calculated an APFS for 10^6 random gene groups, each random group chosen uniformly from the PPR dataset to match the size of the original gene group. The fraction of random groups with a larger APFS than the original group yielded an empirical p-value for the original group's APFS. Of the 1311 clusters, 502 had a significant APFS [$P \leq 0.05$; false-discovery rate (FDR) = 5.3%] (see Figure 1).

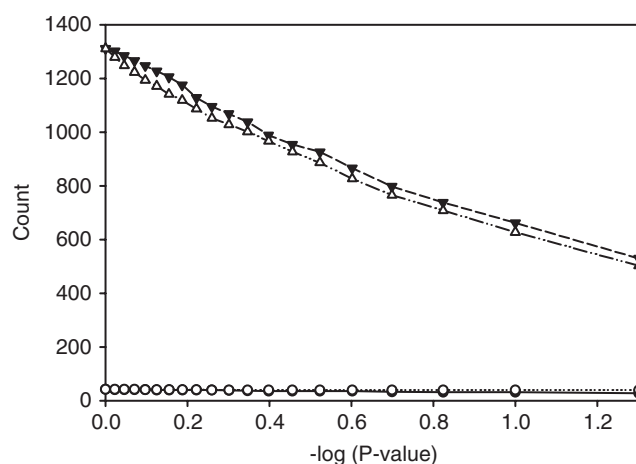


Figure 1. Empirical P -values of clusters estimated from simulation. The figure plots the count of clusters whose empirical P -value did not exceed a particular threshold against the P -value threshold. The empirical P -values were estimated from microarray data (closed triangles) and GO-derived functional similarity data (open triangles). The empirical P -values of nonconserved clusters are shown separately for the microarray data (closed circles) and GO-derived functional similarity data (open circles).

Many clusters correspond to co-expressed gene groups

If a (statistically significant) cluster represents TFBS instances with a common function, the corresponding gene group might be co-expressed. Accordingly, we analyzed expression patterns in microarray experiments from the GNF Atlas 2 (19). The microarray Atlas facilitated the generation of a cross-table, where the rows correspond to 7914 genes downstream of the PPRs and the columns to normalized expression values for 74 human tissues. Consider two clusters and the two corresponding gene groups. For each gene group, the cross-table yielded a pairwise Pearson correlation coefficient (PCC) for their expression values. The substitution of the PCC for the APFS in the procedure above yielded an empirical PCC P -value. Of the 1311 clusters, 529 had a statistically significant PCC ($P \leq 0.05$; FDR = 2.6%) (Figure 1). As further validation of biological functionality, 273 words had both a significant APFS (functional) and a significant PCC (co-expression) similarity ($P \leq 0.05$).

Having established that the gene groups tended to have common GO functions or co-expression, we then examined their tissue-specificity. In a particular tissue (column), the cross-table from the microarray Atlas implicitly ranks each gene (row) according to its expression. For each gene group, the Mann–Whitney rank sum statistic quantifies the expression enrichment for a particular tissue in the gene group relative to other genes. Among the $1311 \times 74 = 97\,014$ gene group-tissue pairs, 1737 showed enriched expression (at $P \leq 0.001$ without multiple test-correction, corresponding to a FDR 5.58%). Of the 1311 gene groups, 450 showed enrichment in at least one tissue, with 58 groups showing enrichment in more than 10 tissues. The vast majority of the gene groups (about 525 of the 1311 groups) showed enriched expression specifically in white blood cells (dendritic, NK, B and T cells),

Table 1. Tissue specificity of DNA words and their association with known transcription factors

DNA Word	Factor	Enriched Tissues	<i>P</i> -value (expression similarity)	<i>P</i> -value (GO functional similarity)
CCGGAAGC	Sp1, c-Ets-1, Ets-1, GABP-alpha, GABP-beta, STAT1, STAT3	PBCD4+Tcells, PBCD8+Tcells, Prostate	1.00E-06	3.24E-03
CGCGATGG	Egr-1	Adrenal gland	1.00E-06	1.49E-02
GCCGCCAT	YY-1	PBCD4+Tcells, PBCD8+Tcells	1.00E-06	4.30E-05
GCCTGCGC	NRF1, Sp1, Sp3	Thyroid	1.00E-06	8.88E-03
GGCGGGGC	Sp1, Sp3, NF-Y	Amygdala, Prostate	1.00E-06	7.42E-03
GGTCACGT	Sp1, Sp3, ATF-1	PLACENTA	1.00E-06	1.81E-02
TTCCGCGC	E2F1, Sp3	PBCD4+Tcells, PBCD8+Tcells, Thymus	2.20E-02	2.24E-02

The last columns give the *P*-values of clusters (estimated by simulation from microarray data and GO-derived functional similarity data). The rows in the table reflect the lexicographic order of the words in column 1.

Table 2. TFBS words corresponding to two clusters, and thereby displaying possible positional regulation of TF tissue specificity

DNA Word	Factor	Distance from TSS (bp)	Tissues	Activation(+)/ Repression(-)
CAGGTGAG	ER-alpha, T3R-beta1, Sp1, Ets	158, ^a 104 ^b	WHOLEBLOOD, ^a Amygdala ^b	+
CGCCCCGC	E2F-1, AP-2alphaA, NRF-1, Egr-1	-59, ^a -176 ^b	Cardiac Myocytes, ^a Uterus Corpus ^b	-
CGCCGCCG	AP-2alphaA, c-Ets-2, Sp1	14, ^a -17 ^b	BM-CD71+EarlyErythroid, ^a Thyroid, ^a fetalbrain ^b	+
GCGGCGGG	p53	20, ^a -56 ^b	TrigeminalGanglion, ^a Appendix ^b	-
GCGGGGCC	Sp1, Sp3, MyoD, AP-2beta	-57, ^a -15 ^b	Bronchialepithelialcells, ^a 721_B_lymphoblasts, ^b SmoothMuscle ^b	+
GGCGGCGC	Sp1, HIF-1, GKLF, NF-Y, CTCF	-39, ^a 19 ^b	Ovary, ^a AdrenalCortex, ^b Appendix, ^b OlfactoryBulb ^b	-

For simplicity, only three examples have been provided for each case (activation and repression). The rows in the table reflect the lexicographic order of the words in column 1. Each word corresponds to TFs in column 2. Each word corresponds to two clusters, whose average positions relative to the TSS are in column 3. Superscripts link the entries in columns 3 and 4, to indicate the relation between the position-specific binding and the tissue-specific regulation of each TF.

generally agreeing with the conclusion of a recent study on motif discovery in the human genome (28).

Recall that 273 gene groups had both common GO functions and co-expression at significant levels ($P \leq 0.05$). Of the corresponding 273 words, 114 exactly matched subsequences of known human TFBSs in TRANSFAC. Additionally, experimental evidence linked 44 of the TF motifs in TRANSFAC matching positionally significant words to one or more of the tissues showing enrichment of the matched word's gene group. Table 1 lists the predicted tissue of enriched expression and the TRANSFAC TF for a randomly selected subset of these 44 words. Supplementary Table 1 in the additional files gives the complete information for all 44 words.

Positional preference is essential to establishing the trends described above; standard sequence analysis alone is insufficient. Two additional lines of evidence support our hypothesis linking TFBSs' locations with tissue-specific usage. First, if a single word corresponded to two or three different positional clusters, the clusters often corresponded to gene groups expressed in strikingly different tissues. For instance, the TRANSFAC binding element for AP-2alphaA, c-Ets-2, Sp1, represented by consensus CGCCGCCG, yields two significant clusters at -17 and +14bp, respectively. While the upstream cluster showed overexpression in fetal brain, the downstream cluster showed overexpression in BM-CD71+EarlyErythroid and Thyroid. Thus, these TFs might use position-specific binding to drive differential tissue-specific activation.

Other factors exhibiting a similar phenomenon include ER-alpha, T3R-beta1, Sp1, Ets (CAGGTGAG) and Sp1, Sp3, MyoD, AP-2beta (GCGGGGCC). On the other hand, some factors might use position-specific binding to cause tissue-specific repression. Such factors include p53 (GCGGCGGG), Sp1, HIF-1, GKLF, NF-Y, CTCF (GGCGGCGC) and Sp1, Sp3, MyoD, AP-2beta (GCGGGGCC) (Table 2). Second, for each of the 273 clusters described earlier, we selected a gene group of equal size as a negative control. Each gene in the control group had a PPR containing the relevant word between -200 and +100bp, but with no other positional restriction. As expected, the Mann-Whitney rank sum test showed that unlike the actual gene groups, the control gene groups did not display any noticeable tissue specificity (see Supplementary Figure S2a and S2b).

The position of a TFBS can influence its function

There were 78 words (about 6.4% of all 1226 words) with two or three different significant clusters. These 78 words presented a unique opportunity to see whether the sequence of a TFBS is sufficient to determine its biological function. The 78 words generated 92 pairs of gene groups, each pair corresponding to a single eight-letter word but to two different positional clusters. If a TFBS had diverse roles (e.g. activation and repression) in different tissues, it might yield a pair of gene groups with significantly different expression patterns across the 74 tissues in the

Table 3. TFBS words corresponding to two clusters, whose gene groups have significantly different microarray expression patterns

DNA Word	Factor	Distance from TSS (bp)	P-value
CCCCGCCC	c-Myc, AP-2alphaA, E2F-1, NF-AT1, MAZ	−65, −20	2.24E-04
CCGCCGCC	YY1, Egr-1, AP-2alphaA, Sp1, Sp3	63, 13	4.18E-02
CGCCCCGC	Sp1, Sp3, E2F-1, Egr-1	−176, 41	4.04E-02
CGCCGCTG	Unidentified	15, 39	4.66E-04
CGGGCGGC	DP-1, E2F:DP, Sp1, GKLf	−15, 23	1.16E-07
GAGGCGGC	Unknown	−16, 20	2.85E-02
GGGCGGCG	Sp1, NF-Y, GKLf	−20, 143	1.15E-11

The rows in the table reflect the lexicographic order of the words in column 1. Each word corresponds to TFs in column 2. Each word corresponds to two clusters, whose average positions relative to the TSS are in column 3. Fisher inverse chi-squared test yielded a (multiple test corrected) two-sided *P*-value (in column 4), which quantifies the overall differences in expression between the gene group pair in the 74 tissues.

GNF Atlas 2 microarray dataset. For each of the 92 pairs, a one-sided Mann–Whitney *P*-value quantified the relative expression of the two gene groups in the 74 tissues. The Fisher inverse chi-squared test (29) assessed the product of the 74 one-sided Mann–Whitney *P*-values, and its two-sided *P*-value for the product indicated the overall differences in expression between the two groups (see Supplementary data—Section 1.4). After multiplying by 92 to correct for multiple testing, seven pairs were statistically significant ($P \leq 0.05$). Table 3 presents results for significant pairs of gene groups ($P \leq 0.05$, after multiplying by 74 to correct for multiple testing).

A comparison of results from positional and sequence-based methods

For a TFBS conserved across several species, comparative genomics uses a multiple alignment across the species to narrow the TFBS search to regions of high conservation (7,30). Positional regulomics might have at least two potential advantages over comparative genomics in identifying TFBSs. First, because positional regulomics does not require accurate sequence alignments, it can find TFBSs in poorly conserved regions. Second, it does not depend on undependable details of the background DNA sequence, thereby reducing the false positive rate of its predictions.

The first potential advantage suggests the following question. Do comparative genomics and its requirement for sequence conservation obscure TFBSs that positional regulomics might find? Let a cluster be partially or less conserved if >20% of positions in it occur in nonconserved regions within the human genome, as determined by human/mouse genome alignment (hg17/mm7 assembly) of the UCSC Genome Browser. Of the 1311 clusters, 42 clusters contained 20% or more positions in nonconserved regions; of these, 12 contained 75% or more positions in nonconserved regions. Thus, sequence conservation considerations had little influence on the 1311 clusters of positions. Out of the 42 nonconserved clusters, 26 and 29 clusters appeared significant ($P < 0.05$) under our analysis using expression and functional similarity data, respectively (Figure 1).

To assess the second potential advantage and to compare false positives from positional and comparative genomics, consider a recent study that identified 54 702 putative human TFBSs by aligning human, mouse,

rat and dog genomes (7). The present study identified 46 670 putative TFBSs, a comparable number. The spatial distribution of transposable elements (TEs) around the TSS may be an indicator of the relative false positive rates in the two studies. TEs comprise about 45% of the human genome and might contribute a substantial fraction of regulatory elements (31,32). However, a sharp decline of TEs around the TSS (33) indicates selection against their insertion in functionally important regions like core promoters where many regulatory elements are positioned. RepeatMasker <<http://www.repeatmasker.org>> was used to determine TE locations using the RepBase library of repeats (34). The total TE count was 24 878, including SINEs, LINEs, LTR elements, DNA elements and other unclassified elements. Overall, the masked regions represented 23% of our dataset.

Figure 2 shows distributions of positions relative to the TSS: Figure 2A, of TE-rich regions; Figure 2B, of ‘comparative TFBSs’ [predicted in (7)]; and Figure 2C, of ‘positional TFBSs’ [predicted in the present study]. TE-rich regions overlapped with 122 comparative TFBSs but with only 50 positional TFBSs (two-sided Fisher exact $P = 7.8 \times 10^{-6}$). Positional TFBSs had a tight distribution from about −200 to +100 bp relative to the TSS, whereas comparative TFBSs were relatively widespread, from about −500 to +500 bp. The positional TFBSs become rare as TEs become common away from the TSS. Figure 2 suggests that the positional methods are relatively insensitive to input sequence lengths, because they predict TFBSs only near their genomic anchor, namely, the TSS in the present study. In any case, Figure 2 suggests that in the cases examined, the putative positional TFBSs contain fewer false positives than the putative comparative TFBSs.

Positional regulomics can identify sets of co-regulating TFBSs and co-regulated genes

TFs combine to form *cis*-regulatory modules (CRMs), complexes controlling gene transcription. Thus, a CRM interacts with certain TFBSs and controls certain genes. The following graphical method predicts co-regulating TFBSs and co-regulated genes, without prior knowledge of the specific TFs in the CRM. By applying the techniques of systems biology to CRMs, the method enhances the dependability and interpretability of predictions.

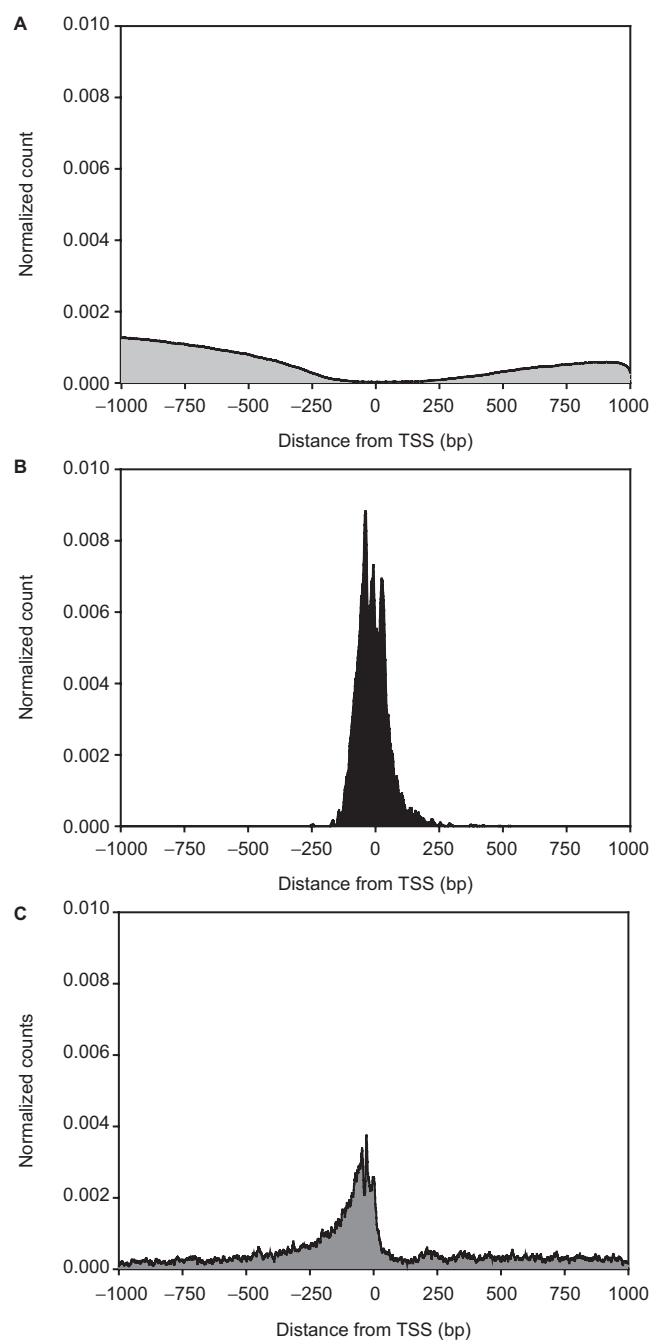


Figure 2. Density of regulatory and repetitive DNA sequences in human core promoters. The plot displays results for 7914 human core promoters. Its X-axis runs from -1000 bp to $+1000$ bp, relative to the TSS for each promoter at 0 bp. The Y-axis represents the normalized count of: (A) TE-derived sequences; (B) TFBSs predicted with our positional methods and (C) TFBSs predicted with phylogenetic footprinting. In each case, the raw counts were normalized to make the area under each graph 1. The boundaries of the three curves indicate the density of predicted sequences in the different regions. Our methods tend to predict TFBSs in the $[-200, +100]$ region of core promoters.

We assembled a dataset containing all genes with complete GO and microarray GNF Atlas 2 data and corresponding to at least one significant cluster. The resulting 3589 genes constituted nodes in each of three networks (i.e. graphs), corresponding to three sources of

information: (i) the positionally significant words, (ii) the GO annotation and (iii) the microarray Atlas. A Pearson correlation coefficient quantified pairwise similarity between genes, based on the significant words occurring in their promoters (see the Methods section). In the corresponding networks, an edge joined a gene pair, if the pair scored above the 95th percentile for the corresponding measure: (i) 0.566 for the Pearson correlation coefficient quantifying TF similarity; (ii) 0.588 for the GO semantic similarity or (iii) 0.546 for the PCC from the microarray Atlas data. The 95th percentile was an arbitrary choice, because considerations of computational time precluded a thorough exploration of possible thresholds.

The three sources of information validated each other's conclusions as follows. In Figure 3A, UPGMA (unweighted pair group method with arithmetic mean) clustered the genes by GO semantic similarity; in Figure 3B, by the similarity of the set of positionally significant words contained in the corresponding promoters. The organized patterns of color in Figure 3 display the correlations between the three sources of information (GO, microarray Atlas data and positional regulomics), so the sources validated each other. Integration of positional, functional and co-expression information generated an intersection network (see the Methods section). Figure 4 shows the gene expression profiles for the most densely connected set of genes, sharing common positional, functional and co-expression properties. Some other profiles appear in Supplementary Figure S4. Therefore, positional regulomics can be combined with (and validated by) other sources of information, to identify modules of TFBSs and coregulated genes.

Algorithm and Datasets

A C++ computer program implemented the algorithm identifying significant clusters of eight-letter words in anchored promoter sequences. A UNIX-compatible version of the program with user-tunable parameters is available for download at the following URL: ftp://ftp.ncbi.nlm.nih.gov/pub/marino/published/positional_regulomics/, along with the pairwise GO functional similarities for 3589 transcripts.

DISCUSSION

Historically, the lambda repressor was the first experimental system known to us to show that position (as well as sequence) influences a TFBS's function. Using the TSS as a genomic landmark, positional regulomics provides strong statistical evidence that in human transcription, the phenomenon is not isolated: if not commonly, at least not rarely, a TFBS's position as well as its sequence can influence the strength of activation or repression of a gene. Some TFs (e.g. AP-2alphaA, ER-alpha, Sp1, Sp3, p53, NRF-1) appear to bind to different positions relative to the TSS, to regulate different genes in different tissues. Moreover, a TFBS's position appears to influence biological function, not just strength of that function. These conclusions rely on data about exact words

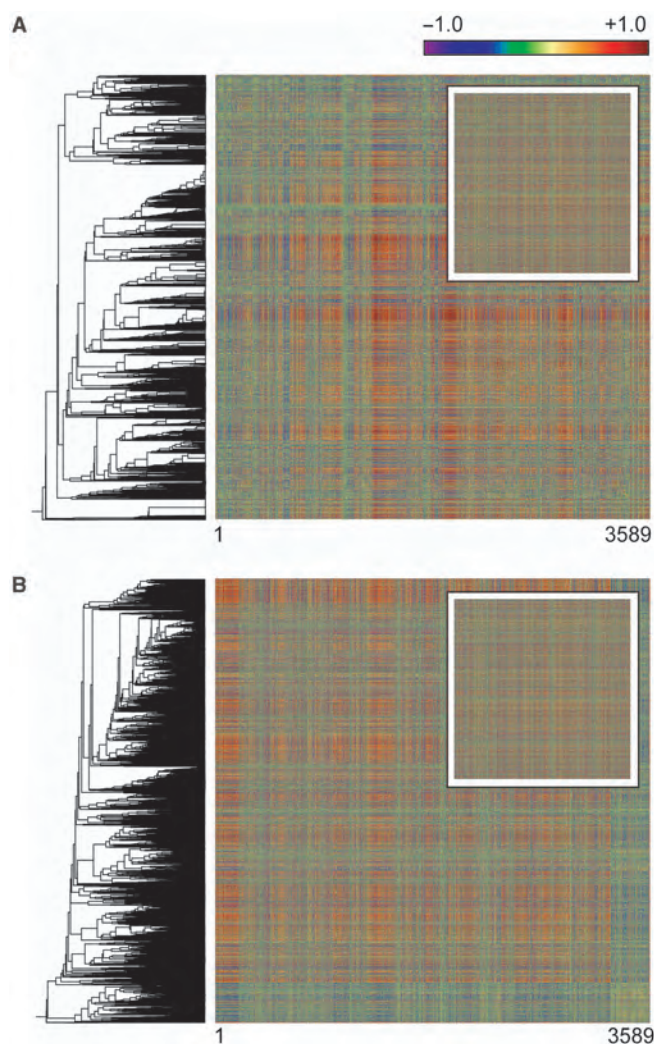


Figure 3. Gene profile correlation matrices. The UPGMA method clustered genes by their GO-derived functional similarity. The matrix in Figure 3A orders the genes identically on both axes by functional similarity. Each off-diagonal element in the matrix corresponds to a pair of different genes. The color of an element codes the Pearson correlation coefficient for the co-expression of the corresponding gene-pair in the microarray data. The off-diagonal blocks of consistent color indicate that functionally similar groups of genes have similar expression patterns. For comparison, the inset in the plot shows a negative control. The inset's matrix orders the genes identically on both axes, but randomly. Accordingly, the matrix lacks off-diagonal blocks of consistent color. The matrix in Figure 3B orders the genes identically on both axes, according to the similarity of the set of positionally significant words contained in the corresponding promoters (see the Methods section). The off-diagonal blocks of consistent color indicate that positional regulomics predicted groups of genes with similar expression patterns.

(i.e. a single sequence pattern with no alternatives), so an analysis based on sequence alone, without position, has no obvious opportunity to draw similar conclusions.

Some experimental results for specific TFBSs support our conclusions about position. The word CCGCCGCC matches the TRANSFAC motif for the YY1 factor and clusters at two different locations (+13 and +63 bp relative to the TSS) (Table 3). The cluster at +63 bp contains transcripts significantly overexpressed in T cells

(PB-CD4+Tcells, PB-CD8+Tcells). In contrast, the cluster located at +13 bp contains transcripts significantly underexpressed in medulla oblongata (Medulla Oblongata). In fact, experimentally, YY1 acts as an activator or repressor, depending on its binding context within a promoter (35,36). Moreover, YY1 enhances transcription in T cells but represses it elsewhere (37). In addition to YY1, our predictions concerning the dual regulatory roles of several other TFs, notably Sp-1 (38), Sp3 (39), and AP-2alphaA (40) matched evidence from experimental literature.

Despite its interesting strengths, our study has some limitations, particularly with respect to alternative promoters. Our dataset contained PPRs corresponding to as many as 4603 genes with putative alternative promoters. In each of these genes, the alternative TSS were spaced at least 500 bp apart (17). Typically, data about functional similarity and microarray expression do not specify possible alternative start sites: the basic unit in both types of data is usually the gene. Alternative promoter usage can have tissue and sequence-context specificity, so the lack of information about alternative promoters probably restricted the precision and scope of our conclusions. If a complete catalogue of annotated promoters and alternative transcripts were available, however, a microarray could use probes with transcript-specific 5' ends to distinguish among alternative promoters. Similarly, GO annotation could distinguish alternative promoters, if it contained the relevant additional information.

In this study, most positions in most clusters were in conserved regions relative to the mouse genome. Because the positions likely represent TFBSs with a common functionality in the human, most such TFBSs likely represent functionality common to both human and mouse. Our methods could not judge, however, the conservation of individual TFBSs in the two genomes or the TFBSs missed (41–44) by phylogenetic analysis (7,30). Variation of individual TFBSs might be one process differentiating species, but our results suggest that only relatively small subsets of TFBSs with a common function display nucleotide changes between human and mouse.

Finally, exact words yield a limited representation of TFBSs. Position-specific scoring matrices (PSSMs) are much more flexible. We are currently implementing improvements to A-GLAM (11), our Gibbs sampler program for finding TFBSs, to combine sequence information with positional information from datasets with genomic anchors, e.g. the TSS. Initial results indicate that position can contribute substantially to the accuracy of sequence motif predictions. Genomic landmarks serve as a 'poor man's alignment', even when precise sequence alignment is impossible. For genes that contain a common TFBS, suggesting co-regulation, our results indicate that positional regulomics can detect positional regulation and thereby unravel the mechanisms underlying diverse functionality and/or expression patterns, by exploiting the location of the TFBS. Further, the resulting models from positional regulomics systematically identify additional genes regulated in a similar manner. Thus, given the success of comparative genomics and its basis in

7. Xie, X.H., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
8. Zhang, C., Xuan, Z., Otto, S., Hover, J.R., McCorkle, S.R., Mandel, G. and Zhang, M.Q. (2006) A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Res.*, **34**, 2238–2246.
9. FitzGerald, P.C., Shlyakhtenko, A., Mir, A.A. and Vinson, C. (2004) Clustering of DNA sequences in human promoters. *Genome Res.*, **14**, 1562–1574.
10. Marino-Ramirez, L., Jordan, I.K. and Landsman, D. (2006) Multiple independent evolutionary solutions to core histone gene regulation. *Genome Biol.*, **7**, R122.
11. Tharakan, K., Marino-Ramirez, L., Sheetlin, S., Landsman, D. and Spouge, J.L. (2005) Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics*, **21**, 1440–1448.
12. Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F. *et al.* (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
13. Suzuki, Y., Yamashita, R., Nakai, K. and Sugano, S. (2002) DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
14. Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K. *et al.* (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.*, **36**, 40–45.
15. Marino-Ramirez, L., Spouge, J.L., Kanga, G.C. and Landsman, D. (2004) Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res.*, **32**, 949–958.
16. Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
17. Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H. *et al.* (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.*, **16**, 55–65.
18. Azuaje, F., Wang, H. and Bodenreider, O. (2005) In ISCB (ed.), *Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies*, Detroit, MI, The International Society for Computational Biology, San Diego, CA, pp. 9–10.
19. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
20. Maglott, D.R., Katz, K.S., Sicotte, H. and Pruitt, K.D. (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, **28**, 126–128.
21. Stormo, G.D. and Fields, D.S. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, **23**, 109–113.
22. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
23. Bader, G.D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.
24. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
25. Al-Shahrour, F., Diaz-Uriarte, R. and Dopazo, J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
26. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
27. Hvidsten, T.R., Laegreid, A. and Komorowski, J. (2003) Learning rule-based models of biological process from gene expression time profiles using gene ontology. *Bioinformatics*, **19**, 1116–1123.
28. Vardhanabhuti, S., Wang, J. and Hannenhalli, S. (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res.*, **35**, 3203–3213.
29. Bailey, T.L. and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
30. Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A. and Johnston, M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
31. Jordan, I.K., Rogozin, I.B., Glazko, G.V. and Koonin, E.V. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.*, **19**, 68–72.
32. Marino-Ramirez, L. and Jordan, I.K. (2006) Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol. Direct*, **1**, 20.
33. Marino-Ramirez, L., Lewis, K.C., Landsman, D. and Jordan, I.K. (2005) Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet. Genome Res.*, **110**, 333–341.
34. Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, **16**, 418–420.
35. Shi, Y., Seto, E., Chang, L.S. and Shenk, T. (1991) Transcriptional repression by YY1, a human GLI-Kruppel-related protein, and relief of repression by adenovirus E1A protein. *Cell*, **67**, 377–388.
36. Yang, W.M., Inouye, C., Zeng, Y., Bearss, D. and Seto, E. (1996) Transcriptional repression by YY1 is mediated by interaction with a mammalian homolog of the yeast global regulator RPD3. *Proc. Natl Acad. Sci. USA*, **93**, 12845–12850.
37. Ji, H.B., Gupta, A., Okamoto, S., Blum, M.D., Tan, L., Goldring, M.B., Lacy, E., Roy, A.L. and Terhorst, C. (2002) T cell-specific expression of the murine CD3delta promoter. *J. Biol. Chem.*, **277**, 47898–47906.
38. Innocente, S.A. and Lee, J.M. (2005) p53 is a NF-Y- and p21-independent, Sp1-dependent repressor of cyclin B1 transcription. *FEBS Lett.*, **579**, 1001–1007.
39. Ammanamanchi, S., Freeman, J.W. and Brattain, M.G. (2003) Acetylated sp3 is a transcriptional activator. *J. Biol. Chem.*, **278**, 35775–35780.
40. Rietveld, L.E., Koonen-Reemst, A.M., Sussenbach, J.S. and Holthuis, P.E. (1999) Dual role for transcription factor AP-2 in the regulation of the major fetal promoter P3 of the gene for human insulin-like growth factor II. *Biochem. J.*, **338**(Pt 3), 799–806.
41. O'Lone, R., Frith, M.C., Karlsson, E.K. and Hansen, U. (2004) Genomic targets of nuclear estrogen receptors. *Mol. Endocrinol.*, **18**, 1859–1875.
42. Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C. *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.
43. Roh, T.Y., Wei, G., Farrell, C.M. and Zhao, K. (2007) Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res.*, **17**, 74–81.
44. Alkema, W. and Wasserman, W.W. (2003) Understanding the language of gene regulation. *Genome Biol.*, **4**, 327.

Widespread Positive Selection in Synonymous Sites of Mammalian Genes

Alissa M. Resch,* Liran Carmel,* Leonardo Mariño-Ramírez,* Aleksey Y. Ogurtsov,* Svetlana A. Shabalina,* Igor B. Rogozin,* and Eugene V. Koonin*

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

Evolution of protein sequences is largely governed by purifying selection, with a small fraction of proteins evolving under positive selection. The evolution at synonymous positions in protein-coding genes is not nearly as well understood, with the extent and types of selection remaining, largely, unclear. A statistical test to identify purifying and positive selection at synonymous sites in protein-coding genes was developed. The method compares the rate of evolution at synonymous sites (K_s) to that in intron sequences of the same gene after sampling the aligned intron sequences to mimic the statistical properties of coding sequences. We detected purifying selection at synonymous sites in $\sim 28\%$ of the 1,562 analyzed orthologous genes from mouse and rat, and positive selection in $\sim 12\%$ of the genes. Thus, the fraction of genes with readily detectable positive selection at synonymous sites is much greater than the fraction of genes with comparable positive selection at nonsynonymous sites, i.e., at the level of the protein sequence. Unlike other genes, the genes with positive selection at synonymous sites showed no correlation between K_s and the rate of evolution in nonsynonymous sites (K_a), indicating that evolution of synonymous sites under positive selection is decoupled from protein evolution. The genes with purifying selection at synonymous sites showed significant anticorrelation between K_s and expression level and breadth, indicating that highly expressed genes evolve slowly. The genes with positive selection at synonymous sites showed the opposite trend, i.e., highly expressed genes had, on average, higher K_s . For the genes with positive selection at synonymous sites, a significantly lower mRNA stability is predicted compared to the genes with negative selection. Thus, mRNA destabilization could be an important factor driving positive selection in nonsynonymous sites, probably, through regulation of expression at the level of mRNA degradation and, possibly, also translation rate. So, unexpectedly, we found that positive selection at synonymous sites of mammalian genes is substantially more common than positive selection at the level of protein sequences. Positive selection at synonymous sites might act through mRNA destabilization affecting mRNA levels and translation.

Introduction

It is well established that nonsynonymous sites in protein-coding sequences are subject to purifying selection caused by constraints operating at the level of protein structure and function and that positive selection that, at least in mammals, affects a minority of genes and/or sites is an important force of adaptive evolution (Li 1997; Vallender and Lahn 2004; Bustamante et al. 2005; Nielsen et al. 2005). Synonymous (silent) sites are often used as a proxy for neutral evolution. Under this premise, the traditional gauge of selection in nonsynonymous sites is the ratio of nonsynonymous (K_a) over synonymous (K_s) substitutions. $K_a/K_s < 1$ is thought to indicate purifying selection, whereas $K_a/K_s > 1$ is construed as the signature of positive selection (Li 1997; Hurst 2002). However, the neutrality of synonymous substitutions is only a rough and not necessarily valid approximation; the extent, range, and underlying causes of selection in synonymous sites remain subjects of intense debate (Chamary, Parmley, and Hurst 2006). The results of several studies suggest that efficient translation (Ikemura 1985; Akashi and Eyre-Walker 1996; Eyre-Walker and Keightley 1999) and mRNA stability (Duan and Antezana 2003; Chamary and Hurst 2005; Shabalina, Ogurtsov, and Spiridonov 2006) are substantial forces of purifying selection in synonymous sites. It has also been shown that synonymous substitutions are under purifying selection in mammalian exonic splicing enhancer motifs (ESEs) (Yeo et al. 2004; Parmley, Chamary, and Hurst

2006) and in alternatively spliced exons (Xing and Lee 2005).

By contrast, to the best of our knowledge, positive selection in synonymous sites has not been reported. However, this possibility has been brought up in the course of analysis of the SPANX family of mammalian cancer-testis genes (Kouprina et al. 2004) and the insect cyclin A inhibitor gene (Avedisov et al. 2001) that are characterized by exceptionally high and comparable rates of evolution in both synonymous and nonsynonymous sites.

We were interested in addressing the problem at a fundamental level: is positive selection in synonymous positions a common phenomenon, and if so, what could be the underlying causes of such selection? We reasoned that, to investigate selection in synonymous sites, the substitution rate in intronic sequences (K_i) was a logical choice of the proxy for neutral evolution. A method for estimating the neutral rate using K_i has recently been reported (Hoffman and Birney 2007). In principle, at least, cases of negative selection in synonymous sites were identified as $K_s/K_i < 1$ whereas cases of positive selection were indicated by $K_s/K_i > 1$. No part of the genome can be automatically assumed to evolve neutrally: the possibility of a hidden function that constrains evolution or an adaptive component in the evolution of a sequence always should be considered. However, apart from pseudogenes, internal regions of introns are among the best candidates for neutrally evolving sequences. The sequences of ~ 30 nucleotides at each end of an intron are thought to be subject to weak purifying selection that stems from the involvement of these sequences in splicing (Louie, Ott, and Majewski 2003; Yeo et al. 2004) and SAS (unpublished observations). In addition, some of the introns contain highly conserved sequences with various, often unknown functions including genes for noncoding RNAs (Washietl et al. 2005). However, in

Key words: synonymous sites, nonsynonymous sites, positive selection, purifying selection, introns.

E-mail: koonin@ncbi.nlm.nih.gov; rogozin@ncbi.nlm.nih.gov.

Mol. Biol. Evol. 24(8):1821–1831. 2007

doi:10.1093/molbev/msm100

Advance Access publication May 23, 2007

Published by Oxford University Press 2007.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

mammals, these functional regions have been estimated to comprise <5% of the intronic sequences (Waterston et al. 2002). In addition, it has been demonstrated that, even in conserved noncoding sequences such as those found in introns, the pressure of purifying selection tends to be substantially weaker than in coding regions (Kryukov, Schmidt, and Sunyaev 2005). Thus, it appears that, after discarding terminal regions, introns could serve as a reasonable approximation of neutrally evolving sequences.

We found that *Ks* and *Ki* distributions for mammalian genes have different statistical properties, which makes *Ki* suspect as the neutral baseline for the analysis of selection in synonymous sites. Therefore we developed a computational procedure to shuffle aligned intron sequences such that their statistical properties mimic those of nonsynonymous sites and used the corresponding substitution rate (*Ki*-pseudo) to assess the extent of negative (purifying) and positive (diversifying) selection in synonymous sites of mammalian genes. It is shown that both types of selection in synonymous sites are widespread and that positive selection at synonymous sites is much more common than positive selection on protein sequence. Positive selection at synonymous sites is unrelated to functional constraints at the protein level, but is linked to gene expression, probably through mRNA destabilization.

Materials and Methods

Identification of Orthologous Genes

We calculated rates of divergence in coding and noncoding DNA for mouse-rat orthologs taken from the May 2004 HomoloGene database (Wheeler et al. 2006). HomoloGene orthologs are defined as bidirectional best hits using the BLAST program for sequence comparison (Altschul et al. 1997). Protein and mRNA sequences were obtained from the Entrez protein and nucleotide databases (Wheeler et al. 2006). We started with a total 8,178 mouse-rat orthologs but removed over half (see below) to eliminate the potential bias estimates of the *Ks* estimates that could be introduced by alternative splice variants and other alignment ambiguities (see next section). The final gene set employed for all analyses contained 1,562 mouse-rat orthologs.

Coding and Noncoding DNA Alignments

Protein alignments for mouse and rat were generated using the MUSCLE alignment package (Edgar 2004). Protein alignments were then used to guide alignment of the corresponding mouse and rat coding sequences (CDS). It was required that each coding sequence contain a start and stop codon, in order to eliminate all partial sequences. All alignments that contained insertions/deletions with the total length >30 bp were removed in order to exclude potential effects of incorrect gene prediction and alternative splicing.

Intron alignments were generated using the OWEN alignment tool (Ogurtsov et al. 2002) with the following specifications: (1) an intron must be bound on 5'/3' ends by exons that align across $\geq 80\%$ of length, (2) the presence

of constitutive splice sites at each intron/exon boundary was required, (3) a *P* value <0.001 for each intron alignment was required, (4) 30-nucleotide regions from the 5'/3' ends of each intron were removed, and 5) the proximal, 5'-terminal introns in the compared orthologous genes were discarded, because these introns are known to be enriched for various regulatory elements and, consequently, could be subject to purifying selection (Majewski and Ott 2002). These requirements help ensure that accurate orthologous intron alignments are generated. The 30-nucleotide regions from the ends of each alignment were removed to eliminate splicing signals from the estimates of intron divergence.

Comparison of Substitution Rates in Coding and Intronic Sequences

The evolutionary rates for coding DNA were originally calculated using the Pamilo-Bianchi-Li method (Li 1993; Pamilo and Bianchi 1993), which takes into account transition and transversion rates. Evolutionary rates for noncoding DNA were measured using the Kimura's 2-parameter model (Kimura 1980). However, considering the difference in the statistical properties of the CDS and intron sequences (see Results and Discussion), a method was developed to shuffle the intron sequence alignments such that their statistical properties mimicked those of coding sequences. Mouse-rat pseudo-CDS alignments were generated from alignments of mouse and rat intronic sequences using the following procedure for each alignment (supplementary fig. 1S): (1) Start the pseudo-CDS from ATG for both mouse and rat sequences. (2) Take the next pentanucleotide starting from the first codon position from the mouse CDS sequence and find this pentanucleotide in the mouse intronic sequence; if there are several such pentanucleotides, 1 is chosen randomly. (3) Add the corresponding segment of the intronic alignment to the pseudo-CDS; if the length of the pseudo-CDS alignment >5 nucleotides, the overlapping 2 nucleotides are chosen randomly; if a pentanucleotide is not found in the mouse intronic sequence, the corresponding fragment of the CDS alignment is added to the pseudo-CDS alignment. (4) The procedure is repeated until the end of the CDS alignment is reached. The resulting pseudo-CDS alignment has the same length as the CDS alignment, and the base compositions of mouse CDS and pseudo-CDS are identical. The significance of the difference in the codon composition of the rat CDS and pseudo-CDS was tested using a Monte-Carlo modification of the χ^2 test (Adams and Skopek 1987).

Detecting Positive and Negative Selection in Synonymous Sites

For each CDS alignment, 10,000 pseudo-CDS alignments were generated. A score of divergence at synonymous sites *Ks* was calculated using the Pamilo-Bianchi-Li method (Li 1993; Pamilo and Bianchi 1993) or the fraction of mismatches at 4-fold degenerate sites. This score was calculated for the mouse-rat CDS alignment (*Ks*) and 10,000 pseudo-CDS alignments (*Ki*-pseudo). The distribution of *Ki*-pseudo was used to calculate

probabilities $P(K_s \leq K_i\text{-pseudo})$ and $P(K_s \geq K_i\text{-pseudo})$; the fractions of the pseudo-CDS with $K_s \leq K_i\text{-pseudo}$ and $K_s \geq K_i\text{-pseudo}$ were taken as approximations of the respective P values. The genes with $P(K_s \geq K_i\text{-pseudo}) \leq 0.05$ were considered positively selected, and the genes with $P(K_s \leq K_i\text{-pseudo}) \leq 0.05$ were considered negatively selected. These calculations were performed for the complete alignments and repeated after masking CG, TG, and CA dinucleotides. For the analysis of statistical properties of distributions and correlation analysis, a pseudo-CDS alignment was randomly drawn from the total sample of 10,000, and $K_i\text{-pseudo}$ was calculated using the PBL method.

Microarray Expression Analysis

The GNF Gene Expression Atlas2 data (Su et al. 2004) for mouse was used as the source of data on genes (rat expression data was limited for the majority of genes and therefore was not included in this analysis). The GNF Atlas2 data contain 2 replicates for each of 61 mouse tissues. The data for redundant tissue types was combined to yield a final set of 55 mouse tissues. Average expression values for each probe were calculated using raw expression data. Average probe expression values for raw data were calculated by summing the expression values across each probe set and dividing that sum by the total number of tissues (55). Tissue breadth values for each gene (the number of tissues that each probe is expressed in) were obtained using raw expression data. The raw expression value for a given tissue had to be ≥ 350 in order for that tissue to be counted in the tissue breadth analysis (Jordan, Marino-Ramirez, and Koonin 2005). Thus, the final tissue breadth score for a probe represents the number of all tissues with the raw expression value ≥ 350 .

Codon Usage

The effective number of codons (ENC) for the coding sequences in the analyzed gene sets was calculated using previously described methods (Wright 1990). The codon adaptation index (CAI) scores (Sharp and Li 1987) for the analyzed coding sequences were calculated using the EMBOSS bioinformatics suite (Rice, Longden, and Bleasby 2000). The CAI values were calculated by comparing the codon usage patterns of a given gene against the codon usage patterns of a reference set of highly expressed mouse genes. Specifically, the GNF Atlas2 mouse expression data were used to identify the top $\sim 10\%$ (1,479/15,007) most highly expressed mouse genes by calculating the average overall expression level of each probe from raw expression data. The average expression values were ranked, from largest to smallest, to obtain the top 10%.

Distance Between Distributions

In order to quantify the dissimilarities between the distribution functions of K_a , K_s , K_i , and $K_i\text{-pseudo}$, we have computed pairwise distances between these distributions using an information-theoretic measure known as the

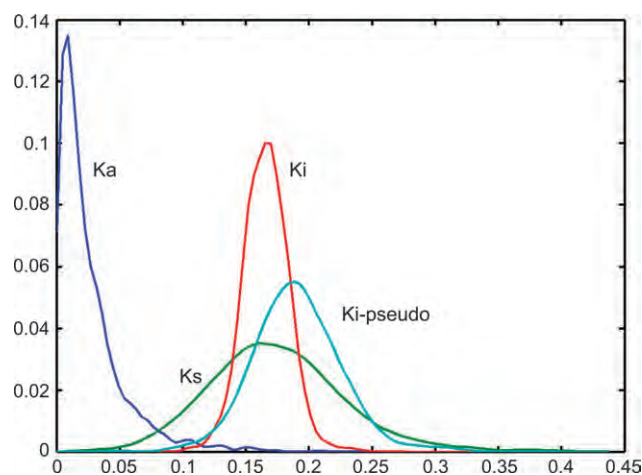


FIG. 1.—The distributions of K_a , K_s , K_i , and $K_i\text{-pseudo}$ in the analyzed set of 1,562 rodent genes.

L-divergence (Lin 1991). This distance measure is a refined version of the widely used Kullback-Leibler distance.

Results and Discussion

Using Intron Evolution Rate as the Baseline for Detecting Selection in Synonymous Sites

In order to avoid ambiguities of alignment, especially, in intron sequences, as well as substitution saturation effects, we limited the present analysis to orthologous genes from closely related rodents, mouse and rat. It has been recently shown that K_i is particularly prone to taxon-specific variation at longer evolutionary distances (Hoffman and Birney 2007). A critical issue is whether K_s/K_i is an adequate measure of selection in synonymous sites. We generated K_s and K_i distributions for a set of 1,562 reliable (see Materials and Methods) alignments of intronic and coding sequences from orthologous mouse and rat genes in order to assess the suitability of K_i as the baseline for detecting selection in synonymous sites. First, we compared the statistical properties of the distributions of K_i and K_s . The distribution of K_i had almost precisely the same mean and median as the K_s distribution but was quite narrow compared to the latter: the standard deviation of K_s was more than twice greater than that of K_i (fig. 1 and supplementary table 1S). Furthermore, the skewness of the distribution was also much greater for the K_s distribution than for the K_i distribution (fig. 1 and supplementary table 1S). We also compared the nucleotide compositions of introns and synonymous sites and found substantial differences between these two (supplementary table 2S). These observations showed that K_s and K_i distributions had distinct statistical properties and suggested that introns and synonymous positions in exons are subject to different evolutionary forces.

Previous studies that compared K_s and K_i do not seem to arrive to a consensus. Some reports have claimed that synonymous substitution rates are approximately equal to those in introns despite differences in the patterns of substitution (Hughes and Yeager 1997; Chamary and Hurst

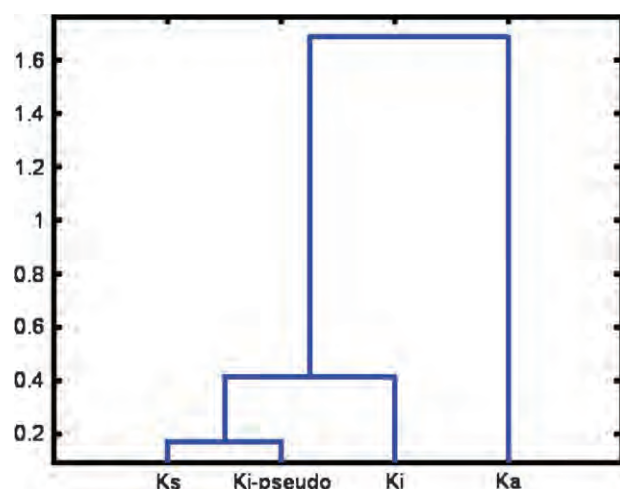


FIG. 2.—A dendrogram based on the pairwise distances between the distributions of Ka, Ks, Ki, and Ki-pseudo. The distances between the distributions are in bits.

2004), whereas others have suggested that intron rates of divergence are greater than those in synonymous sites (Hellmann et al. 2003), or conversely, that synonymous substitution rates exceed those found in introns (Subramanian and Kumar 2003). Our present observation that Ks and Ki are (nearly) identical on average but are very differently distributed suggests that these diverging conclusions might be attributed to different evolutionary models and data sets used in the respective studies. It has been argued that the apparent increase in synonymous substitution rates of some genes over those of introns is due to the context-dependence of mutation in synonymous sites, in particular, the high mutation rate of CpG dinucleotides (Hughes and Yeager 1997; Kondrashov, Ogurtsov, and Kondrashov 2006).

Our observations, together with those in previous studies, suggest that Ki might not be a proper null model for Ks due to different nucleotide compositions of coding and non-coding DNA and distinct statistical properties of the Ks and Ki distributions. Thus, we developed a computational procedure to account for these differences between introns and synonymous sites. Under this approach, alignments of pseudo-coding sequence (pseudo-CDS) were generated by sampling alignments of intronic sequences such as to mimic the base composition of the synonymous sites for each respective gene and thus eliminate potential artifacts caused by differences in the CpG content and other compositional differences between synonymous sites and introns. The pseudo-CDS alignments were used to calculate Ki-pseudo values (see Material and Methods and Supplementary fig. 1S for details). Using an information-theoretical measure of divergence (see Materials and Methods for details), we computed distances between the distributions and found that, unlike Ki, Ki-pseudo had statistical properties highly similar to those of Ks (fig. 2 and supplementary table 3S). In particular, the distribution of Ki-pseudo was shifted to the right compared to the Ki distribution such that the right tail of the Ki-pseudo distribution behaved more similarly to that of the Ks distribution

Table 1
The Number of Genes with Significant Positive and Negative Selection in Synonymous Sites

Method	Positive Selection	<i>P</i>	Negative Selection	<i>P</i>
1) PBL, all sites	188	4×10^{-26}	517	$<10^{-50}$
2) PBL, CG, TG, CA removed	175	6×10^{-21}	279	$<10^{-50}$
3) PBL, CX, XG (X = A,T,G,C) removed	185	5×10^{-25}	218	6×10^{-40}
4) 4F, all sites	189	10^{-27}	438	$<10^{-50}$
5) 4F, CG, TG, CA removed	151	2×10^{-12}	227	2×10^{-44}

NOTE.—Selection in synonymous sites was measured using the Pamilo-Bianchi-Li (PBL) method and the fraction of mismatches at 4-fold degenerate sites (4F). The probability of finding this many or more cases of apparent positive or negative selection by chance was calculated using the binomial test; the expected number of genes in the positive and the negative set each is 78 (5% of 1,562 genes). The dramatic loss of sites caused by the CX/XG masking procedure made the 4F analysis (unlike the PBL method) inapplicable for this method. Therefore, the results obtained with this stringent filtering were not used for constructing the final sets of genes with apparent negative and positive selection in synonymous sites.

(fig. 1). Accordingly, Ki-pseudo values were used as the baseline to detect purifying and positive selection acting on synonymous sites of mammalian genes.

Partitioning Rodent Genes into Negatively Selected, Neutral, and Positively Selected Sets Using Synonymous Sites As the Criterion

We examined positive and negative selection in synonymous sites, using the Ks/Ki-pseudo ratio as the criterion under 2 distinct estimation schemes, the Pamilo-Bianchi-Li (PBL) method (Li 1993; Pamilo and Bianchi 1993) and the fraction of mismatches at 4-fold degenerate sites (4F method). These calculations were performed either before or after masking CG, TG, and CA dinucleotides (the highly mutable CpG sites and the “highly CpG-prone” sites, i.e., those convertible to CpG via a single transition [Kondrashov, Ogurtsov, and Kondrashov 2006]) or, finally, after removing all CpX and XpG dinucleotides (all CpG-prone sites). Starting with a set of 1,562 reliably aligned mouse-rat orthologs (see Materials and Methods), we identified a significant excess (compared to the random expectation) of genes with both negative and positive selection in synonymous sites in all 5 tests. Masking the mutable dinucleotides did not substantially affect the results (table 1). In order to obtain conservative estimates of positively and negatively selected genes, we required agreement between the 2 evolutionary models: only genes found to be positively or negatively selected in 3 or 4 tests were included in the final sets. The results of test #3 (table 1, PBL, CpX, and XpG sites removed) were not used in this selection procedure because of the dramatic loss of sites ($>50\%$) that was caused by the masking procedure and made the 4F method inapplicable. However, the results of the PBL test show that this masking had but a small effect on the number of genes with apparent negative and positive selection in synonymous sites (table 1).

With this approach, 185 cases of positive (diversifying) selection (positive set) and 438 cases of negative (purifying) selection (negative set) were identified. The

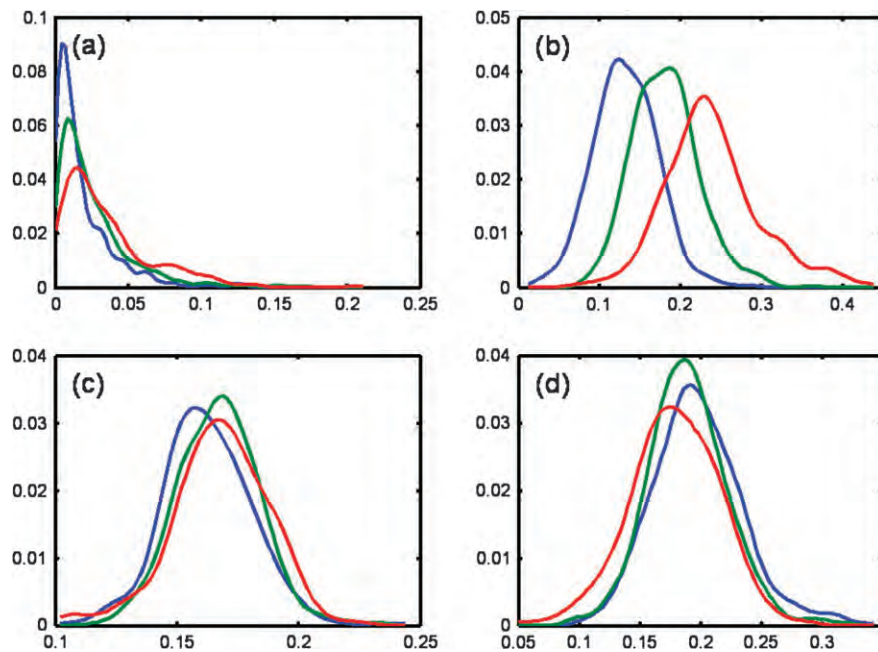


FIG. 3.—Distributions of K_a (a), K_s (b), K_i (c), and K_i -pseudo (d) for the 3 sets of rodent genes. In each panel, the blue curve corresponds to the negative set, the green curve to the neutral set, and the red curve to the positive set.

remaining 939 genes were conservatively assigned to the neutral set. Thus, $\sim 28\%$ of the analyzed rodent genes were found to be subject to substantive negative selection in synonymous sites, whereas roughly half as many genes ($\sim 12\%$) appeared to be subject to relatively strong positive selection. A comparison of the content of CpG sites involving synonymous position in the 3 gene sets did not reveal significant differences, suggesting that the observed distinct modes of evolution are not caused by effects of mutagenic contexts (supplementary table 4S).

We further compared the K_s , K_i , K_i -pseudo, and K_a distributions for the positive, negative, and neutral datasets. The distributions of K_a , K_i , and K_i -pseudo were very similar, in both shape and parameter values in all 3 sets (figs. 3a,c,d), although there was a statistically significant difference between the distributions of all 3 of these variables (supplementary table 5S). In a sharp contrast, the K_s distributions differed much more significantly between the 3 gene sets than the K_a , K_i -pseudo, or K_i distributions (see supplementary table 5S, and compare fig. 3b to figs. 3a,c,d). Although all 3 K_s distributions were, approximately, equally broad, the negative and positive set distributions were significantly shifted to the left and to the right, respectively, compared to the neutral set distribution (fig. 3). In particular, the mean K_s in the positive set was substantially greater than the mean K_s of the negative set (mean = 0.239 for positive; mean = 0.132 for negative set) as expected of sites evolving under positive selection.

It should be emphasized that the existence of a major difference between the K_s distributions but not K_i distributions for the negative and positive sets (compare fig. 3b and fig. 3c) all but rules out a potential alternative explanation of these results, namely, that the apparent positive selection in synonymous sites is an artifact caused by an anomalously high sequence conservation, due to purifying selection, in

the intronic sequences of the respective genes. In order to further ascertain that purifying selection in introns was not a significant factor accounting for the detected positive selection in synonymous sites, we applied stringent filtering to remove potential functional elements from the intronic sequences used in the K_i and K_i -pseudo calculations. For that purpose, longer exon-flanking sequences were trimmed off the intron alignment, and short introns that could be enriched for functional elements were discarded. This procedure reduced the set of orthologous gene pairs available for the analysis to 952 but did not substantially change the fractions of genes subject to positive and negative selection in synonymous sites (table 6S). These results indicate that K_i -pseudo is, indeed, an appropriate baseline for measuring selection acting at other classes of sites in orthologous genes from closely related species.

K_s and K_a Are Correlated in the Negative and Neutral Sets but Not in the Positive Set

Does the relationship between K_s , K_i -pseudo, and K_a reveal anything about the evolutionary forces that affect the positive and negative sets? We addressed this question by checking whether any of the variables were correlated, and whether the strength of such correlations differed between the sets. We observed a moderate but statistically highly significant positive correlation between K_s and K_a for the negative set (table 2 and fig. 4), which could be expected, given that genes in this set are under strong purifying selection; similar observations have been reported previously in several independent studies (Lipman and Wilbur 1984; Wolfe and Sharp 1993; Mouchiroud, Gautier, and Bernardi 1995; Makalowski and Boguski 1998; Smith and Hurst 1999). A somewhat weaker but also highly significant correlation between K_a and K_s was seen in the

Table 2
The Correlations Between the Analyzed Variables

Negative Set							
	Ka	Ks	Ki	Ki-pseudo	EL	EB	ΔG
Ka	1.00	0.27	0.29	0.13	-0.10	-0.20	0.13
	Ks	1.00	0.46	0.67	-0.14	-0.15	-0.11
		Ki	1.00	0.55	-0.08	-0.18	0.09
			Ki-pseudo	1.00	-0.08	-0.11	-0.12
				EL	1.00	0.69	0.13
					EB	1.00	0.09
						ΔG	1.00
Neutral Set							
	Ka	Ks	Ki	Ki-pseudo	EL	EB	ΔG
Ka	1.00	0.19	0.24	0.14	-0.08	-0.21	0.09
	Ks	1.00	0.53	0.67	-0.06	-0.08	0.05
		Ki	1.00	0.58	0.03	-0.03	0.08
			Ki-pseudo	1.00	0.01	-0.03	-0.01
				EL	1.00	0.66	0.06
					EB	1.00	-0.02
						ΔG	1.00
Positive Set							
	Ka	Ks	Ki	Ki-pseudo	EL	EB	ΔG
Ka	1.00	0.08	0.02	-0.01	-0.17	-0.25	0.25
	Ks	1.00	0.48	0.72	0.15	0.18	0.17
		Ki	1.00	0.54	0.10	-0.04	0.14
			Ki-pseudo	1.00	0.16	0.09	0.05
				EL	1.00	0.53	-0.02
					EB	1.00	-0.20
						ΔG	1.00
Complete Set							
	Ka	Ks	Ki	Ki-pseudo	EL	EB	ΔG
Ka	1.00	0.25	0.23	0.08	-0.10	-0.22	0.14
	Ks	1.00	0.46	0.44	-0.05	-0.07	0.10
		Ki	1.00	0.54	0.01	-0.08	0.10
			Ki-pseudo	1.00	0.01	-0.03	-0.06
				EL	1.00	0.65	0.07
					EB	1.00	-0.01
						ΔG	1.00

NOTE.—The table shows the Pearson correlation coefficient (r) values for each pair of variables. Statistically significant ($P \leq 0.05$) values are indicated by shading.

neutral set (table 2 and fig. 4). The significant correlation between K_s and K_a in the negative and neutral sets implies that the evolutionary forces that exert purifying selection on synonymous sites in the negative set are linked to the evolution of protein structure and function. In particular, it seems likely that the negative selection acting on synonymous sites has to do with the high level of expression that is characteristic of genes encoding highly conserved proteins (Pal, Papp, and Hurst 2001; Krylov et al. 2003; Wolf, Carmel, and Koonin 2006). By contrast, in the positive set, K_s and K_a showed a much weaker and not significant correlation ($r = 0.08$, $P = 0.28$; table 2 and fig. 4). To control for the possible effect of the smaller sample size of the positive set, we generated 10,000 random samples of 185 genes each from the total dataset and found only 93 sampled sets with $r \leq 0.08$ ($P < 0.01$). Thus, the absence of a significant correlation between K_a and K_s in the positive gene set is not a sample size artifact. This observation suggests that, in sharp contrast with both the negative and the neutral sets, the forces that affect the evolution of synonymous sites in the positive-set genes are uncoupled from any selection acting at the level of protein structure and function.

We then examined the relationship between K_s and K_i -pseudo and observed strong positive correlations for

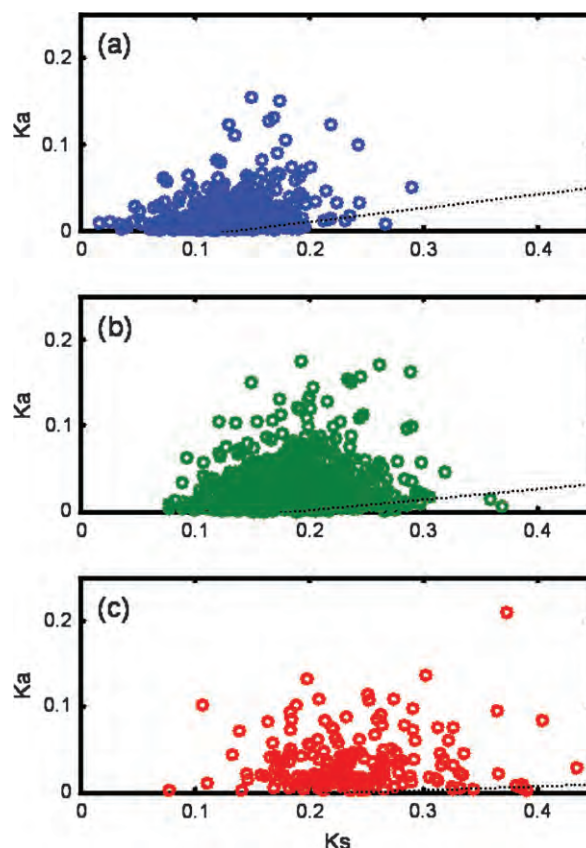


FIG. 4.—The correlations between K_a and K_s for the negative (a), neutral (b), and positive (c) gene sets.

all 3 gene sets; slightly weaker but also highly significant correlations were found for K_s and K_i (table 2). At first glance, this result suggests the possibility that evolution of introns might not be neutral, and accordingly, K_i -pseudo might not be a robust null model for measuring selection at synonymous sites. However, this does not seem to be the case, because the strength of the correlation was nearly identical among the 3 gene sets. It appears most likely that the correlations between K_s and K_i (and K_i -pseudo) reflect regional mutational biases across the genome. Such biases have been reported previously (Matassi, Sharp, and Gautier 1999), and for the rodent genes analyzed here, we observed a highly significant anticorrelation between the differential of the K_i and K_s values and the distance separating the respective genes on the chromosome: closely spaced genes, typically, had similar K_s , K_i , and K_i -pseudo values; no such effect was seen for K_a (supplementary figs. 2S and 3S).

Given that K_s is correlated with K_a in the negative and neutral sets, and with K_i in all 3 sets, we performed a partial correlation analysis in an attempt to disentangle these correlations. In the negative and neutral sets, the correlations between K_a and both K_s and K_i became smaller but remained highly significant after the removal of the effect of the other variable (supplementary table 7S). Thus, in the negative and neutral sets, the correlation between K_a and K_s appears to be valid in itself and might reflect similar selective pressures at synonymous and nonsynonymous sites. The correlation between K_a and K_i , which was, in

Table 3
Codon Bias (ENC and CAI), Gene Expression (EL and EB), and mRNA Stability (ΔG) in the 3 Sets of Analyzed Rodent Genes

		Positive		Neutral		Negative	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Codon Bias	ENC	49.618	4.245	49.890	4.081	49.199	4.400
	CAI	0.748	0.046	0.754	0.044	0.759	0.045
Gene Expression	EL	464.450	752.336	470.100	595.625	508.509	605.628
	EB	17.219	19.126	18.232	20.491	19.583	20.429
mRNA Stability	ΔG	-0.329	0.044	-0.338	0.042	-0.347	0.045

NOTE.— ΔG values are normalized for length.

part, independent of the Ks-Ki correlation and was greater in the negative set than in the neutral set (supplementary table 7S), is harder to explain. It cannot be ruled out that there is some pressure of purifying selection on intron sequences, the nature of which remains obscure. Should such a selective component, indeed, affect evolution of introns in the negative set, this would make our estimate of genes subject to purifying selection at synonymous sites even more conservative. By contrast, in the positive set, there was no significant correlation between Ka and either Ks or Ki, indicating that, in these genes, evolution of the protein sequence is completely decoupled from the evolution of noncoding sequences. Taken together, these results indicate that there are at least 2 distinct components in the evolution of synonymous sites, a selective one and a mutational one. The nature of the mutational component is the same across all analyzed genes. By contrast, the selective component is linked to the protein evolution in the negative set but apparently is of a different nature in the positive set.

Potential Driving Forces of Selection in Synonymous Sites: Significant Differences in Expression and mRNA Stability Between the Positive and Negative Sets

What factors contribute to positive and negative selection in synonymous sites? Perhaps even more importantly, can we identify the probable causes of the lack of correlation between Ks and Ka in the positive set? We evaluated the roles of gene function, codon bias, gene expression, and mRNA stability as potential driving forces of selection in synonymous sites. There were no significant differences in the distribution across the Gene Ontology (GO) categories between the genes of the negative, neutral, and positive sets (data not shown), hence no straightforward explanation for the observed differences in the selection regimes through the biological functions of the respective genes. We then compared the codon bias [determined either as the effective number of codons (ENC) or as the codon adaptation index (CAI)] between the 3 gene sets. Moderate but statistically significant differences in CAI were detected between the negative and positive sets, with the highest value observed for the negative set (tables 3 and 4). Thus, the pattern of codon bias exhibited by the genes in the negative set is more similar to the pattern found among highly expressed genes (the reference set) than to the pattern found within the pos-

Table 4
Statistics of the Comparison of the Negative, Neutral, and Positive Gene Sets with Respect to Codon Bias (ENC and CAI), Gene Expression (EL and EB) and mRNA Stability (ΔG)

	Codon Bias		Gene Expression		mRNA Stability ΔG
	ENC	CAI	EL	EB	
Positive versus Neutral*	1.000	0.259	1.000	1.000	0.0117
Positive versus Negative*	0.818	0.012	1.000	0.747	2.7×10^{-16}
Negative versus Neutral*	0.013	0.115	1.000	1.000	3.96×10^{-10}
Combined**	0.017	0.009	0.63	0.468	3.69×10^{-18}

NOTE.—Bonferroni adjusted *P* values computed using Student's *t*-test (*) and *P* values for combined data were computed using ANOVA (**).

itive set. This result is consistent with the expectation that genes that are more biased in their choice of synonymous codons tend to be more conserved. A significant difference in ENC was also observed between the negative and neutral sets (tables 3 and 4). Given that a tighter control over codon usage would be a side effect of strong purifying selection within the negative set, this result is not surprising. The nonsignificant differences between the positive set and the other 2 sets are likely to result from the fast evolution in synonymous sites of positively selected genes. A comparison of gene expression, determined either as expression level (EL) or as expression breadth (EB), revealed no significant differences between the positive, neutral, and negative sets (table 3 and 4). Some reports have suggested that purifying selection on synonymous sites affects the efficiency and accuracy of translation in certain model organisms such as *Escherichia coli* and *Drosophila melanogaster* (Ikemura 1985; Eyre-Walker 1996; Akashi and Eyre-Walker 1998); however, no strong indications of such selective pressures in mammalian genomes have been detected (Smith and Hurst 1999; Duret and Mouchiroud 2000; Iida and Akashi 2000). Furthermore, these findings were in agreement with previous reports indicating that rates of synonymous divergence are not correlated with patterns in gene expression (Lercher, Chamary, and Hurst 2004).

However, examination of the correlations between the rates of evolution in synonymous and nonsynonymous sites and characteristics of expression in the 3 gene sets produced more informative and, partly, unexpected results. In the negative set, there was a relatively low but statistically significant anticorrelation between Ks and expression (both EL and EB); these anticorrelations paralleled those between Ka and expression (table 2) and were compatible with the previous observations on slow evolution of highly and broadly expressed genes (Duret and Mouchiroud 2000; Pal, Papp, and Hurst 2001; Krylov et al. 2003; Wolf, Carmel, and Koonin 2006). The positive set showed a strikingly different pattern, with Ks being positively correlated with both EL and EB, whereas for Ka the correlation was negative and of roughly the same magnitude as in the other 2 gene sets (table 2). Thus, in a pattern that is the diametrical opposite of what is seen in the negative set (and, less pronouncedly, in the neutral set), fast evolution in synonymous sites that appear to be subject to positive selection is associated with higher and broader expression of the corresponding gene.

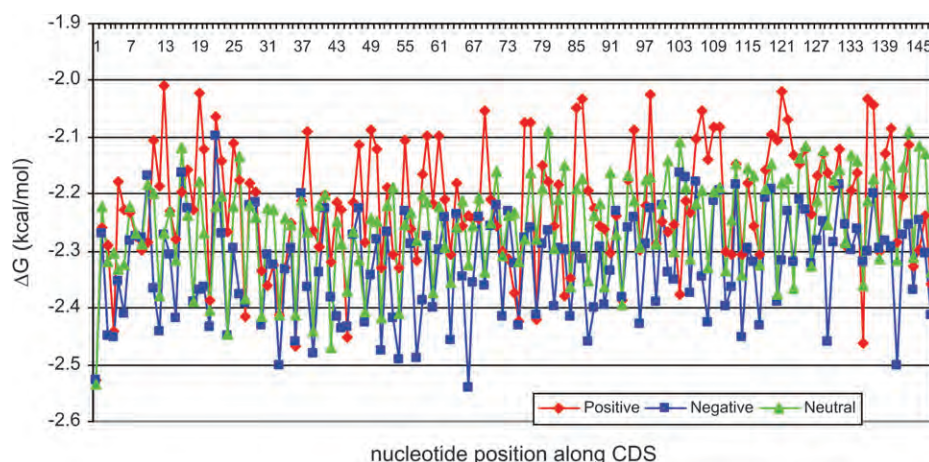


FIG. 5.—Plot of ΔG values (kcal/mol) calculated for base pairs along the 150-nucleotide stretch of coding sequence starting from the codon immediately following the start ATG codon. Values are averaged across the CDS in the negative set (blue), neutral set (green), and positive set (red).

It has been proposed that purifying selection on synonymous sites is linked to increased mRNA stability (Duan and Antezana 2003; Chamary and Hurst 2005; Shabalina, Ogurtsov, and Spiridonov 2006). Thus, we looked for differences in patterns of mRNA stability between the positive, neutral, and negative sets. Using previously published methods (Shabalina, Ogurtsov, and Spiridonov 2006), we found that the average predicted mRNA stability (kcal/mol) was significantly greater in the negative set than in the neutral or positive sets (fig. 5 and tables 5 and 6). It has been shown previously that the contributions of nucleotides to mRNA stability followed a periodic pattern dictated by the structure of the genetic code, with the third, degenerate position being the primary contributor (Shabalina, Ogurtsov, and Spiridonov 2006). This pattern was, indeed, apparent in all 3 gene sets analyzed here (fig. 5). Notably, the difference in RNA stability (estimated free energy) between the negative, neutral, and positive sets was consistent and significant over all 3 codon positions (fig. 5 and tables 5 and 6). This suggests that positive selection for mRNA destabilization affects all codon positions, although in nonsynonymous sites this relatively weak effect is overshadowed by selection acting at the protein level. No significant differences in nucleotide content within the codon positions were observed between the positive and negative sets (supplementary table 8S), indicating that the differences in mRNA stability are not artifacts caused by different base compositions.

We further assessed correlations between ΔG and K_s in the 3 gene sets and found the results to be consistent with selection acting to maintain or establish the optimal stability of the mRNA secondary structure. The correlation between ΔG and K_s was not significant in the neutral set, whereas the correlations of opposite signs were observed in the negative and the positive sets. In the negative set, there was a low but significant anticorrelation between ΔG and K_s , whereas the positive set showed a somewhat greater, positive correlation (table 2). In other words, in the negative set, the genes that evolve relatively slowly tend to possess less stable mRNA secondary structure than faster evolving genes; conversely, in the positive set the faster evolving

genes are less stable than slowly evolving ones. Thus, it appears that purifying selection in the negative set prevents the formation of excessively stable secondary structure, whereas in the positive set diversifying selection might drive mRNA destabilization.

To summarize, the correlations between K_s , gene expression, and predicted mRNA stability were all negative in the negative set but positive in the positive set (fig. 6A). These coherent but contrasting correlation structures, certainly, do not prove a cause-and-effect relationship between mRNA stability and expression in mammalian evolution, but are compatible with the hypothesis that both purifying and positive selection in synonymous sites act at the level of mRNA secondary structure, which affects stability and, through mRNA degradation process, the expression levels measured in microarray experiments. In a sharp contrast, the correlation structures for K_a were identical for the positive and negative set (fig. 6B), suggesting similar patterns of (purifying) selection in nonsynonymous sites. In this case, acceleration of evolution, on average, seems to result in mRNA destabilization and lower expression.

Conclusions

We developed and applied a robust statistical test to identify purifying and positive selection acting on synonymous sites of mammalian genes using shuffled intron

Table 5
Free Energy (ΔG) of Base-Pairing for Individual Codon Positions for the 3 Gene Sets

	Positive		Neutral		Negative	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
pos_1	-2.182	0.092	-2.205	0.043	-2.275	0.066
pos_2	-2.175	0.089	-2.197	0.054	-2.276	0.063
pos_3	-2.293	0.091	-2.351	0.051	-2.411	0.06

NOTE.— ΔG values are normalized for the predicted number of base-paired nucleotides for each of the codon positions within the 150 upstream nucleotides of the CDS (AUG start codon removed).

Table 6
Statistics of the Comparison of the Base-Pairing Free
Energies (ΔG) for Individual Codon Positions in the 3
Analyzed Gene Sets

	pos_1	pos_2	pos_3	Total CDS
Positive versus				
Neutral*	0.387	0.456	9.3×10^{-4}	0.0117
Positive versus				
Negative*	7.38×10^{-7}	3.6×10^{-8}	3.51×10^{-10}	2.7×10^{-16}
Negative versus				
Neutral*	1.19×10^{-7}	1.0×10^{-8}	2.95×10^{-6}	3.96×10^{-10}
Combined**	3.54×10^{-9}	1.45×10^{-10}	1.09×10^{-12}	3.69×10^{-18}

NOTE.—Bonferroni adjusted P values were computed for the same data included in table 5 using Student's t -test (*) and P values for combined data were computed using ANOVA (**).

sequences as the proxy for neutral evolution. As expected, considering many previous reports on strong, positive correlations between K_a and K_s , we observed that a substantial fraction of the analyzed genes was subject to significant purifying selection at synonymous sites (Akashi 1994; Mouchiroud, Gautier, and Bernardi 1995; Makalowski and Boguski 1998). By contrast, the finding that $\sim 12\%$ of the genes seemed to experience substantial positive selection at synonymous sites was surprising. A comparison of the distributions of K_s and K_i for the negative and positive gene sets (fig. 3) seems to rule out the possibility that the apparent positive selection in synonymous sites is actually due to purifying selection affecting the respective intronic sequences.

The fraction of rodent genes with apparent positive selection in synonymous codons is considerably greater than the fraction of mammalian genes that have been reported to exhibit positive selection on the level of the amino acid sequence, as reflected in $K_a/K_s > 1$ for the entire coding sequence. The specific numbers of mammalian genes that are subject to strong positive selection differ between studies, but a recent careful analysis of $\sim 8,000$ orthologous genes from humans and chimpanzees revealed only 35 genes with statistically significant positive selection manifest at the level of the entire protein sequence (Nielsen et al. 2005). Of the 1,562 genes analyzed here, only 2 had $K_a/K_s > 1$ at the statistically significant level, indicating positive selection on the level of complete protein sequences (data not

shown). Thus, it seems that, in comparable tests, 1 to 2 orders of magnitude more mammalian genes exhibit positive selection in synonymous than in nonsynonymous sites. This excess of positive selection in synonymous sites seems to reflect different balances of evolutionary forces that act at positions coding for amino acids and at synonymous positions. Protein sequences are almost universally subject to purifying selection of varying strength ($K_a/K_s \ll 1$ for most genes), which is likely to obscure more subtle effects of positive selection acting at some positions; indeed, site-specific positive selection detected by multiple alignment analysis appears to be common (Yang and Bielawski 2000; Zhang, Nielsen, and Yang 2005). By contrast, as shown here, readily detectable purifying selection on synonymous sites might affect only about one-quarter of the mammalian protein-coding genes such that positive selection is readily detectable against the, largely, neutral background of synonymous sites. Additionally, K_i and, especially, K_i -pseudo appear to be better neutral baselines than K_s such that positive selection at synonymous sites could be easier to detect than positive selection at nonsynonymous sites for which K_s is used as the baseline. This being said, a comparison of the distributions of K_a and K_s in rodent genes (fig. 1) indicates that K_a/K_s remains a reasonable measure of the strength of selection in proteins, given that protein sequences are subject to much more pronounced constraints than noncoding sites.

Finding the biological basis or at least a strong functional correlate of positive selection in synonymous sites turned out to be a challenge. The lack of significant correlation between K_s and K_a in the positive set indicates that positive selection in synonymous sites is decoupled from the evolution of the respective proteins. This conclusion is compatible also with the lack of any significant excess of a particular class of biological functions among the genes in the positive set. By contrast, analysis of links between selection in synonymous sites and gene expression and mRNA stability revealed nontrivial connections. Although there were no significant differences in the overall expression level or breadth between positively selected, negatively selected and neutral genes, the dependences between evolution rate and expression was remarkably different. In particular, and in contrast to negatively selected genes, among genes with positive selection in synonymous sites, those that are highly and widely expressed appear to

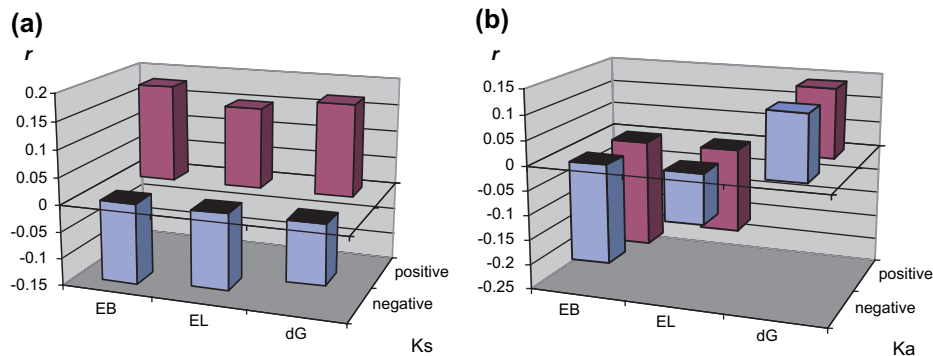


FIG. 6.—The structure of correlations between K_s (A) and K_a (B) and expression breadth (EB), expression level (EL), and predicted mRNA stability (dG) for the positive (purple) and negative (light blue) gene sets.

evolve faster (i.e., under a stronger selective pressure) than lowly expressed ones. The second clear and statistically highly significant correlation was with predicted stability of the mRNA secondary structure: the transcripts in the set of positively selected genes are predicted to be considerably less stable than those in the negative and neutral sets. The correlations between Ks, on the one hand, and expression and mRNA stability, on the other hand, were all of the same sign within the positive and negative sets but of the opposite signs between the sets (fig. 6A). This is compatible with a causal relationship between mRNA stability and expression levels measured in microarray experiments, with the link probably actualized through regulation of mRNA degradation. Therefore, although we technically cannot ascertain the direction of evolution without using a third species as an outgroup, we suspect that mRNA destabilization could be an important factor that, through its effect on mRNA stability and possibly also translation rates, drives positive selection in synonymous sites. Additionally or alternatively, it is conceivable that positive selection in synonymous positions is driven by the necessity to maintain interactions with other RNA species (Mattick and Makunin 2006).

We cannot be confident that the correlates of selection in synonymous site detected here, indeed, reflect the principal underlying selective forces. However, it is our hope that the demonstration of the wide spread of positive selection in synonymous sites in mammalian genes stimulates further theoretical and experimental studies aimed at the deeper characterization of the causes of this phenomenon.

Supplementary Material

Supplementary tables and figures are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

We thank King Jordan, Alexei Kondrashov, Yuri Wolf, and John Wootton for helpful discussions. This work was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS.

Literature Cited

- Adams WT, Skopek TR. 1987. Statistical test for the comparison of samples from mutational spectra. *J Mol Biol.* 194:391–396.
- Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics.* 136:927–935.
- Akashi H, Eyre-Walker A. 1998. Translational selection and molecular evolution. *Curr Opin Genet Dev.* 8:688–693.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.
- Avedisov SN, Rogozin IB, Koonin EV, Thomas BJ. 2001. Rapid evolution of a cyclin A inhibitor gene, roughex, in *Drosophila*. *Mol Biol Evol.* 18:2110–2118.
- Bustamante CD, Fledel-Alon A, Williamson S, et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature.* 437:1153–1157.
- Chamary JV, Hurst LD. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol Biol Evol.* 21:1014–1023.
- Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* 6:R75.
- Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet.* 7:98–108.
- Duan J, Antezana MA. 2003. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. *J Mol Evol.* 57:694–701.
- Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol.* 17:68–74.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
- Eyre-Walker A. 1996. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Mol Biol Evol.* 13:864–872.
- Eyre-Walker A, Keightley PD. 1999. High genomic deleterious mutation rates in hominids. *Nature.* 397:344–347.
- Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B, Paabo S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* 13:831–837.
- Hoffman MM, Birney E. 2007. Estimating the neutral rate of nucleotide substitution using introns. *Mol Biol Evol.* 24:522–531.
- Hughes AL, Yeager M. 1997. Comparative evolutionary rates of introns and exons in murine rodents. *J Mol Evol.* 45:125–130.
- Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18:486.
- Iida K, Akashi H. 2000. A test of translational selection at ‘silent’ sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene.* 261:93–105.
- Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 2:13–34.
- Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. *Gene.* 345:119–126.
- Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16:111–120.
- Kondrashov FA, Ogurtsov AY, Kondrashov AS. 2006. Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. *J Theor Biol.* 240:616–626.
- Kouprina N, Mullokandov M, Rogozin IB, Collins NK, Solomon G, Otstot J, Risinger JI, Koonin EV, Barrett JC, Larionov V. 2004. The SPANX gene family of cancer/testis-specific antigens: rapid evolution and amplification in African great apes and hominids. *Proc Natl Acad Sci USA.* 101:3077–3082.
- Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* 13:2229–2235.
- Kryukov GV, Schmidt S, Sunyaev S. 2005. Small fitness effect of mutations in highly conserved non-coding regions. *Hum Mol Genet.* 14:2221–2229.
- Lercher MJ, Chamary JV, Hurst LD. 2004. Genomic regionalism in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res.* 14:1002–1013.

- Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol.* 36:96–99.
- Li WH. 1997. *Molecular Evolution*. Sunderland, MA: Sinauer.
- Lin J. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans Inform Theory.* 37:145–151.
- Lipman DJ, Wilbur WJ. 1984. Interaction of silent and replacement changes in eukaryotic coding sequences. *J Mol Evol.* 21:161–167.
- Louie E, Ott J, Majewski J. 2003. Nucleotide frequency variation across human genes. *Genome Res.* 13:2594–2601.
- Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* 12:1827–1836.
- Makalowski W, Boguski MS. 1998. Synonymous and non-synonymous substitution distances are correlated in mouse and rat genes. *J Mol Evol.* 47:119–121.
- Matassi G, Sharp PM, Gautier C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr Biol.* 9:786–791.
- Mattick JS, Makunin IV. 2006. Non-coding RNA. *Hum Mol Genet* 15 Spec No. 1:R17–29.
- Mouchiroud D, Gautier C, Bernardi G. 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. *J Mol Evol.* 40:107–113.
- Nielsen R, Bustamante C, Clark AG, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3:e170.
- Ogurtsov AY, Roytberg MA, Shabalina SA, Kondrashov AS. 2002. OWEN: aligning long collinear regions of genomes. *Bioinformatics.* 18:1703–1704.
- Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics.* 158:927–931.
- Pamilo P, Bianchi NO. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol Biol Evol.* 10:271–281.
- Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol.* 23:301–309.
- Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16:276–277.
- Shabalina SA, Ogurtsov AY, Spiridonov NA. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.* 34:2428–2437.
- Sharp PM, Li WH. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15:1281–1295.
- Smith NG, Hurst LD. 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics.* 153:1395–1402.
- Su AI, Wiltshire T, Batalov S, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA.* 101:6062–6067.
- Subramanian S, Kumar S. 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* 13:838–844.
- Vallender EJ, Lahn BT. 2004. Positive selection on the human genome. *Hum Mol Genet* 13 Spec No. 2:R245–254.
- Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol.* 23:1383–1390.
- Waterston R, Lindblad-Toh HK, Birney E, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature.* 420:520–562.
- Wheeler DL, Barrett T, Benson DA, et al. 2006. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 34:D173–180.
- Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. *Proc Biol Sci.* 273:1507–1515.
- Wolfe KH, Sharp PM. 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J Mol Evol.* 37:441–456.
- Wright F. 1990. The ‘effective number of codons’ used in a gene. *Gene.* 87:23–29.
- Xing Y, Lee C. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc Natl Acad Sci USA.* 102:13526–13531.
- Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution.* 15:496–503.
- Yeo G, Hoon S, Venkatesh B, Burge CB. 2004. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci USA.* 101:15700–15705.
- Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* 22:2472–2479.

William Martin, Associate Editor

Accepted May 17, 2007

Origin and Evolution of Human microRNAs From Transposable Elements

Jittima Piriyaopongsa,* Leonardo Mariño-Ramírez[†] and I. King Jordan^{*,1}

^{*}School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332 and [†]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894

Manuscript received February 23, 2007

Accepted for publication April 12, 2007

ABSTRACT

We sought to evaluate the extent of the contribution of transposable elements (TEs) to human microRNA (miRNA) genes along with the evolutionary dynamics of TE-derived human miRNAs. We found 55 experimentally characterized human miRNA genes that are derived from TEs, and these TE-derived miRNAs have the potential to regulate thousands of human genes. Sequence comparisons revealed that TE-derived human miRNAs are less conserved, on average, than non-TE-derived miRNAs. However, there are 18 TE-derived miRNAs that are relatively conserved, and 14 of these are related to the ancient L2 and MIR families. Comparison of miRNA *vs.* mRNA expression patterns for TE-derived miRNAs and their putative target genes showed numerous cases of anti-correlated expression that are consistent with regulation via mRNA degradation. In addition to the known human miRNAs that we show to be derived from TE sequences, we predict an additional 85 novel TE-derived miRNA genes. TE sequences are typically disregarded in genomic surveys for miRNA genes and target sites; this is a mistake. Our results indicate that TEs provide a natural mechanism for the origination miRNAs that can contribute to regulatory divergence between species as well as a rich source for the discovery of as yet unknown miRNA genes.

MICRORNAS (miRNAs) are small, ~22-nt-long, noncoding RNAs that regulate gene expression (AMBROS 2004). In animals, miRNA genes are transcribed into primary miRNAs (pri-miRNAs) and processed by Drosha to yield ~70- to 90-nt pre-miRNA transcripts that form hairpin structures. Mature miRNAs are liberated from these longer hairpin structures by the RNase III enzyme Dicer (BARTEL 2004). Drosha acts in the nucleus, cleaving the pri-miRNA near the base of the hairpin stem to yield the pre-miRNA sequence. The pre-miRNA is then exported to the cytoplasm where the stem is cleaved by Dicer to produce a miRNA duplex. One strand of this duplex is rapidly degraded and only the mature ~22-nt miRNA sequence remains. The mature miRNA associates with the RNA-induced silencing complex (RISC), and together the miRNA-RISC targets mRNAs for regulation. miRNA target specificity is determined by partial complementarity with the 3'-untranslated region (UTR) sequence of the mRNA, and regulation is achieved by translational repression and/or mRNA degradation. miRNAs have been implicated in a variety of functions, including developmental timing (LEE *et al.* 1993; REINHART *et al.* 2000), apoptosis (BRENNER *et al.* 2003), and hematopoietic differentiation (CHEN *et al.* 2004).

miRNAs were first discovered in *Caenorhabditis elegans* through genetic analysis of developmental mutants

(LEE *et al.* 1993). The small RNA product of the *lin-4* gene was found to negatively regulate *lin-14* expression via interaction with a complementary region in the *lin-14* 3'-UTR. This system appeared to be unique until a second example of a similar small regulatory RNA in *C. elegans*, *let-7*, was discovered 7 years later (REINHART *et al.* 2000). Shortly thereafter, *let-7* homologs and transcripts were detected among a phylogenetically diverse set of animals (PASQUINELLI *et al.* 2000). The realization that miRNAs represent a distinct, coherent, and abundant class of regulatory genes was finally crystallized in 2001 with the publication of three back-to-back articles in *Science*, reporting the discovery of numerous novel miRNA genes (LAGOS-QUINTANA *et al.* 2001; LAU *et al.* 2001; LEE and AMBROS 2001). These articles introduced the term miRNA to refer to all small RNAs with similar genomic features but unknown functions, and miRNAs have now been found in all metazoans surveyed for their presence (BARTEL 2004).

Given their relatively recent discovery and characterization, a number of open questions concerning the function and evolution of miRNAs remain. In particular, the evolutionary origins of miRNAs are not well appreciated. For instance, many miRNA genes were found to be evolutionarily conserved and this was thought to be a general characteristic of miRNAs. However, a number of nonconserved miRNAs have been recently discovered (BENTWICH *et al.* 2005). The extent to which miRNA genes evolve as paralogous gene families is also unknown. Even the upper bound on the number of miRNA genes encoded by any given genome is not

¹Corresponding author: School of Biology, Georgia Institute of Technology, 310 Ferst Dr., Atlanta, GA 30332-0230.
E-mail: king.jordan@biology.gatech.edu

known (BEREZIKOV *et al.* 2006), and the number of new entries in the miRBase registry of miRNA genes continues to grow steadily (GRIFFITHS-JONES *et al.* 2006).

We sought to evaluate the contribution of transposable elements (TEs) to the origin and evolution of human miRNA genes. Another class of regulatory RNAs, small interfering RNAs (siRNAs), are known to be related to TEs. Interestingly, this has been pointed out as a distinction between miRNAs and siRNAs, which are closely related in terms of structure, function, and biogenesis. As opposed to siRNAs, miRNAs were thought to derive from loci distinct from other genes or TEs (BARTEL 2004). However, several examples of miRNA genes that are derived from TEs have been recently identified (SMALHEISER and TORVIK 2005; BORCHERT *et al.* 2006; PIRIYAPONGSA and JORDAN 2007). We wanted to look at this phenomenon more closely to identify the full extent of human miRNA genes that are related to TEs and to characterize how these genes evolve as well as their regulatory and functional potential.

TEs have several characteristics that make them interesting candidates for donating miRNA sequences. First of all, TEs are ubiquitous and abundant genomic sequences. Thus, they could provide for the emergence of paralogous miRNA gene families as well as multiple target sites dispersed throughout the genome. Since TEs tend to be among the most rapidly evolving of all genomic sequences, they may also provide a mechanism for the emergence of lineage-specific miRNA genes that could exert diversifying regulatory effects. Finally, the full contribution of TEs to miRNA sequences is likely to be underestimated due to ascertainment biases. This is because computational methods aimed at the detection of novel miRNAs tend to purposefully exclude TE sequences (BENTWICH *et al.* 2005; LINDOW and KROGH 2005; NAM *et al.* 2005; LI *et al.* 2006). This is often done for reasons of tractability, but also reflects the widely held notion that TEs are genomic parasites that do not play any functional role for their host species (DOOLITTLE and SAPIENZA 1980; ORGEL and CRICK 1980). However, many studies have identified a variety ways in which TEs have been domesticated (MILLER *et al.* 1992) to provide functions to their hosts (KIDWELL and LISCH 2001). These cases include the donation of coding sequences (VOLFF 2006) as well as numerous instances of TE-derived regulatory sequences (BRITTEN 1996; JORDAN *et al.* 2003; VAN DE LAGEMAAT *et al.* 2003).

To evaluate the contribution of TEs to human miRNAs, we compared the genomic locations of TEs to the locations of experimentally validated human miRNA sequences reported in the miRBase database (GRIFFITHS-JONES *et al.* 2006). The evolutionary dynamics of TE-related miRNAs were evaluated by within- and between-genome sequence comparisons. The potential regulatory and functional significance of TE-derived miRNAs was explored by combining information on miRNA target-site prediction, expression data for miRNA–mRNA pairs, and gene functional annotations. We also sought to discover putative cases of

novel TE-derived miRNA genes in the human genome through *ab initio* prediction.

MATERIALS AND METHODS

Detection: Human miRNA sequences and predicted target sites were taken from version 8.2 of the miRBase database (GRIFFITHS-JONES *et al.* 2006). These data do not include *ab initio* miRNA gene predictions. The UCSC Genome Browser (KENT *et al.* 2002) and Table Browser (KAROLCHIK *et al.* 2004) tools were used to search for miRNA genes collocated with TEs and to compare the evolutionary rates of miRNA genes. Human miRNA sequences were mapped to the hg18 (NCBI build 36.1) version of the human genome sequence and a generic feature format “custom track” was created (available upon request). Genomic locations of the miRNAs were compared to the locations of TEs annotated with the RepeatMasker program (SMIT *et al.* 1996–2004). For this purpose, precomputed RepeatMasker annotations of hg18 were combined with RepeatMasker-determined genomic locations of a set of 96 “conserved” TE families recently added to Repbase (JURKA *et al.* 2005). These conserved consensus sequences correspond to low-copy-number TEs that show anomalously low levels of between-genome orthologous sequence divergence and can be found by searching Repbase (<http://www.girinst.org/>) with the keyword “conserved.”

Sequences of TE-derived miRNAs were compared to the human genome sequence using BLAT (KENT 2002). The criteria used for genome sequence hits were (1) $\geq 80\%$ sequence identity with the query miRNA sequence and (2) the genomic hit region must be $\geq 80\%$ and $\leq 120\%$ of the length of the miRNA query sequence. The latter requirement was used to ensure that long genomic insertions were not identified as putative paralogous miRNAs.

Evolution: Comparative genomic sequence data from the UCSC Genome Browser were used to analyze the relative evolutionary rates of human miRNAs. Evolutionary rates were derived from multiple whole-genome sequence alignments between the human and 16 other vertebrate genomes (KENT *et al.* 2003; BLANCHETTE *et al.* 2004). Human miRNA evolutionary rates were calculated in two ways: (1) by evaluating the number of conserved sites per miRNA and (2) by evaluating the per-site conservation scores of miRNA sequences. Conserved human genome sites were predicted by the phastCons program, which uses a phylogenetic hidden Markov model to calculate the probabilities of sites being either conserved or nonconserved (SIEPEL *et al.* 2005). Conservation scores for human genome sites were also taken from the phastCons analysis of the vertebrate multiple genome sequence alignment, and these scores correspond to the posterior probability that a site is conserved or nonconserved.

Regulation and function: Human miRNA target-site predictions were taken from miRBase, which uses a modified protocol based on the miRanda algorithm (ENRIGHT *et al.* 2003). The locations of target-site sequences in the human genome were compared to the RepeatMasker-based TE annotations. Expression levels for human miRNAs across five tissues (thymus, brain, liver, placenta, and testis) were taken from an oligonucleotide-based microarray study (BARAD *et al.* 2004). Human mRNA expression levels from corresponding mRNA targets were taken from the Novartis SymAtlas data set (SU *et al.* 2004). Corresponding miRNA and mRNA expression profiles were normalized using standard z-score transformation with the program Spotfire (<http://www.spotfire.com>) and compared using the Pearson correlation coefficient. Gene expression data were visualized using the Genesis program (STURN *et al.* 2002).

Gene ontology (GO) analysis (ASHBURNER *et al.* 2000) was done using the GO Tree Machine program (ZHANG *et al.* 2004). GO Tree Machine was used to identify significantly over-represented biological process GO terms from a set of genes predicted to be regulated by a particular miRNA and to plot the location of these GO terms along the GO-directed acyclic graph.

TE-miRNA prediction: TE locations in the human genome were considered together with the output of the program EvoFold, which combines RNA secondary structure prediction with the evaluation of multiple sequence alignments to identify conserved secondary structures (PEDERSEN *et al.* 2006). TE sequences that encode conserved hairpin structures with length ≥ 55 bp, a single terminal loop ≤ 20 bp, and at least six paired bases in the stem region (BENTWICH *et al.* 2005) were chosen for further analysis. For conserved TE-encoded hairpins of < 55 bp that met all other criteria, the predicted secondary structure sequences were extended manually and rechecked for the ability to form hairpin structures using the program RNAfold from the Vienna RNA package (HOFACKER *et al.* 1994). Sequences that were able to encode hairpins ≥ 55 bp after manual extension were chosen for further analysis. The potential for putative TE-derived miRNAs identified in this way to be expressed was evaluated using EST and mRNA data. Our TE-miRNA prediction protocol is represented in supplemental Figure 1 at <http://www.genetics.org/supplemental/>.

RESULTS

Transposable-element-derived miRNAs: miRBase is an online database of miRNA gene sequences and predicted target sites (GRIFFITHS-JONES *et al.* 2006); version 8.2 of miRBase contained 462 human miRNA gene sequences. Of these human miRNA genes, 379 are defined on the basis of experimental information, cloning of mature miRNA sequences for the most part, while 83 are predictions on the basis of sequence similarity with miRNAs that have been experimentally characterized in related species. We mapped these human miRNA genes to the complete genome sequence and compared their locations to the locations of annotated TEs. A total of 68 human miRNA genes share sequences with TEs, and all but 7 of these correspond to miRNAs experimentally characterized from human samples. The absence of *ab initio* miRNA gene predictions in the miRBase data set ensures that we are uncovering *bona fide* TE-miRNA relationships. Of these TE-related miRNAs, 49 are found in intron sequences while 19 are intergenic.

TE-related miRNAs differ in terms of the extent of overlap with TE sequences and the number of distinct TE sequences from which they are derived. For each individual TE-related human miRNA, a schematic in supplemental Figure 2 (at <http://www.genetics.org/supplemental/>) illustrates the identity of all colocated TE sequences along with the extent and position of the TE-miRNA overlap and the relationship between the strand-specific orientation of the TE and the miRNA. The majority (50 of 68) of TE-related miRNAs consist of $> 50\%$ TE-derived positions (Figure 1A), and this figure is likely to be an underestimate since many TE sequences are known to have diverged beyond the ability to be

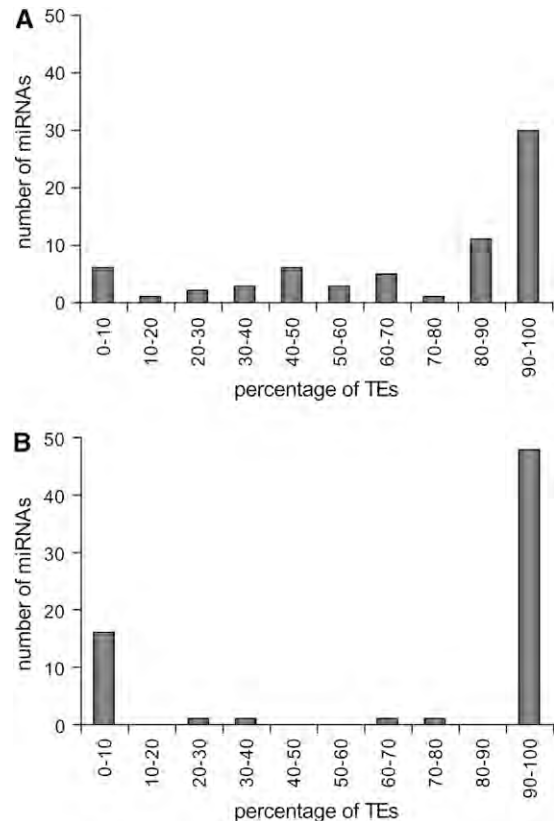


FIGURE 1.—Percentage of TE-derived residues in miRNA genes. Frequency distributions are shown for the percentages of TE-derived residues relative to miRNA gene sequences (A) and mature miRNA sequences (B).

recognized by the RepeatMasker annotation software. The TE-miRNA overlap distribution for the region of the miRNA gene that corresponds to the processed (mature) regulatory sequence is even more bimodal (Figure 1B); 47 sequences have $> 95\%$ of mature miRNA positions covered by TE sequence. Nevertheless, there are a handful (7 of 68) of TE-related miRNA genes that have $< 20\%$ of their sequences colocated with TE sequence. These may represent spurious cases of TE-miRNA overlap. Visual inspection of the TE-miRNA alignments (supplemental Figure 2 at <http://www.genetics.org/supplemental/>) was used to eliminate these unreliable cases. Only the 55 cases with at least 50% TE coverage of the pre-miRNA sequence and/or 100% TE coverage of the mature miRNA sequence were considered as actual TE-derived miRNAs and used for further analysis (Table 1). These 55 TE-derived miRNAs represent $\sim 12\%$ (55/462) of all human miRNAs reported in miRBase version 8.2.

The TE-related miRNAs that we identified are derived from all four major classes of human TEs: long- and short-interspersed nuclear elements (LINE and SINE), long-terminal-repeat-containing elements (LTR) and DNA-type transposons (Table 1). Specific classes and families of TEs show marked over- or underrepresentation among

TABLE 1
TE-derived human miRNAs

miRNA name (from miRBase)	miRBase accession no.	Coordinates ^a	Colocated TE	Overlap ^b	Average conservation score	Targets ^c
hsa-mir-130b	MI0000748	Chromosome 22: 20337593–20337674(+)	MIRm	65.85	0.8492	865 (10.75)
hsa-mir-151	MI0000809	Chromosome 8: 141811845–141811934(–)	L2	100.00	0.9317	863 (12.28)
hsa-mir-28	MI0000086	Chromosome 3: 189889263–189889348(+)	L2	93.02	0.9979	1136 (10.21)
hsa-mir-325	MI0000824	Chromosome X: 76142220–76142317(–)	L2	89.80	0.9905	751 (13.32)
hsa-mir-330	MI0000803	Chromosome 19: 50834092–50834185(–)	MIRm	53.19	0.9867	927 (5.18)
hsa-mir-345	MI0000825	Chromosome 14: 99843949–99844046(+)	MIR	39.80	0.8265	895 (7.82)
hsa-mir-361	MI0000760	Chromosome X: 85045297–85045368(–)	MER5A	81.94	0.9998	882 (14.51)
hsa-mir-370	MI0000778	Chromosome 14: 100447229–100447303(+)	MIRm	100.00	0.9893	1006 (4.77)
hsa-mir-374	MI0000782	Chromosome X: 73423846–73423917(–)	L2	54.17	0.9970	773 (7.50)
hsa-mir-378	MI0000786	Chromosome 5: 149092581–149092646(+)	MIRb	90.91	1.0000	0 (0)
hsa-mir-421	MI0003685	Chromosome X: 73354937–73355021(–)	L2	89.41	0.9999	1023 (14.47)
hsa-mir-422a	MI0001444	Chromosome 15: 61950182–61950271(–)	MIR3	100.00	0.0018	940 (7.34)
hsa-mir-493	MI0003132	Chromosome 14: 100405150–100405238(+)	L2	66.29	0.9990	0 (0)
hsa-mir-513-1	MI0003191	Chromosome X: 146102673–146102801(–)	MER91C	100.00	0.0543	1065 (7.14)
hsa-mir-513-2	MI0003192	Chromosome X: 146115036–146115162(–)	MER91C	100.00	0.0003	1065 (7.14)
hsa-mir-544	MI0003515	Chromosome 14: 100584748–100584838(+)	MER5A1	100.00	0.9337	1056 (10.42)
hsa-mir-545	MI0003516	Chromosome X: 73423664–73423769(–)	L2	82.08	0.9958	1065 (16.345)
hsa-mir-548a-1	MI0003593	Chromosome 6: 18679994–18680090(+)	MADE1	78.35	0.0391	1255 (7.09)
hsa-mir-548a-2	MI0003598	Chromosome 6: 135601991–135602087(+)	LTR16A1, MADE1	100.00	0.0047	1255 (7.09)
hsa-mir-548a-3	MI0003612	Chromosome 8: 105565773–105565869(–)	MLT1G1, MADE1	100.00	0.0044	1255 (7.09)
hsa-mir-548b	MI0003596	Chromosome 6: 119431911–119432007(–)	MADE1	83.51	0.0175	1197 (5.93)
hsa-mir-548c	MI0003630	Chromosome 12: 63302556–63302652(+)	MADE1	83.51	0.0092	1302 (6.76)
hsa-mir-548d-1	MI0003668	Chromosome 8: 124429455–124429551(–)	MADE1	83.51	0.0076	1055 (10.24)
hsa-mir-548d-2	MI0003671	Chromosome 17: 62898067–62898163(–)	MADE1	83.51	0.0000	1055 (10.24)
hsa-mir-552	MI0003557	Chromosome 1: 34907787–34907882(–)	L1MD2	100.00	0.0000	1067 (11.62)
hsa-mir-558	MI0003564	Chromosome 2: 32610724–32610817(+)	MLT1C	45.74	0.0112	778 (7.58)
hsa-mir-562	MI0003568	Chromosome 2: 232745607–232745701(+)	L1MB7	100.00	0.0019	954 (11.64)
hsa-mir-566	MI0003572	Chromosome 3: 50185763–50185856(+)	AluSg	100.00	0.0000	1184 (80.07)
hsa-mir-570	MI0003577	Chromosome 3: 196911452–196911548(+)	MADE1	82.47	0.0000	1115 (4.22)
hsa-mir-571	MI0003578	Chromosome 4: 333946–334041(+)	L1MA9	96.88	0.0000	948 (8.33)
hsa-mir-575	MI0003582	Chromosome 4: 83893514–83893607(–)	MIR	61.70	0.0001	1048 (7.35)
hsa-mir-576	MI0003583	Chromosome 4: 110629303–110629400(+)	L1MB7	100.00	0.0121	921 (10.53)
hsa-mir-578	MI0003585	Chromosome 4: 166526844–166526939(+)	L2	44.79	0.0064	1012 (7.61)
hsa-mir-579	MI0003586	Chromosome 5: 32430241–32430338(–)	MADE1, L1MB8	100.00	0.3543	1202 (6.32)
hsa-mir-582	MI0003589	Chromosome 5: 59035189–59035286(–)	L3, L3	85.71	0.9954	1017 (8.06)
hsa-mir-584	MI0003591	Chromosome 5: 148422069–148422165(–)	MER81	92.78	0.0008	794 (10.96)
hsa-mir-587	MI0003595	Chromosome 6: 107338693–107338788(+)	MER115	100.00	0.0053	970 (6.39)
hsa-mir-588	MI0003597	Chromosome 6: 126847470–126847552(+)	L1MA3	100.00	0.0000	873 (10.77)
hsa-mir-603	MI0003616	Chromosome 10: 24604620–24604716(+)	MADE1	84.54	0.0102	1008 (7.44)
hsa-mir-606	MI0003619	Chromosome 10: 76982222–76982317(+)	L1MCc	100.00	0.0014	776 (8.38)
hsa-mir-607	MI0003620	Chromosome 10: 98578416–98578511(–)	MIR	100.00	0.9990	985 (8.83)
hsa-mir-616	MI0003629	Chromosome 12: 56199213–56199309(–)	L2	100.00	0.0004	922 (10.30)
hsa-mir-619	MI0003633	Chromosome 12: 107754813–107754911(–)	L1MC4, AluSx	100.00	0.0008	765 (8.89)
hsa-mir-625	MI0003639	Chromosome 14: 65007573–65007657(+)	L1MCa	100.00	0.0018	1065 (4.41)
hsa-mir-626	MI0003640	Chromosome 15: 39771075–39771168(+)	L1MB8, L1MCa	56.38	0.0086	1022 (6.65)
hsa-mir-633	MI0003648	Chromosome 17: 58375308–58375405(+)	MIRb	100.00	0.0136	843 (7.12)
hsa-mir-634	MI0003649	Chromosome 17: 62213652–62213748(+)	L1ME3A	48.45	0.0019	886 (5.08)
hsa-mir-640	MI0003655	Chromosome 19: 19406872–19406967(+)	MIRb	100.00	0.0074	853 (28.49)
hsa-mir-644	MI0003659	Chromosome 20: 32517791–32517884(+)	L1MB3	61.70	0.1035	970 (4.95)
hsa-mir-645	MI0003660	Chromosome 20: 48635730–48635823(+)	MER1B	62.77	0.0002	682 (13.49)
hsa-mir-648	MI0003663	Chromosome 22: 16843634–16843727(–)	L2	98.94	0.0008	943 (6.15)

(continued)

TABLE 1
(Continued)

miRNA name (from miRBase)	miRBase accession no.	Coordinates ^a	Colocated TE	Overlap ^b	Average conservation score	Targets ^c
hsa-mir-649	MI0003664	Chromosome 22: 19718465–19718561(–)	L1M4, MER8, AluSx	100.00	0.0005	1033 (10.65)
hsa-mir-652	MI0003667	Chromosome X: 109185213–109185310(+)	MER91C	100.00	0.9883	803 (39.36)
hsa-mir-659	MI0003683	Chromosome 22: 36573631–36573727(–)	Arthur1	46.39	0.0027	890 (8.20)
hsa-mir-95	MI0000097	Chromosome 4: 8057928–8058008(–)	L2	95.06	0.9862	847 (16.06)

^a Human genome (hg 18) coordinates of the miRNA.

^b Percentage of miRNA overlapping with TE sequence.

^c Total number of targets with the percentage derived from TEs in parentheses.

human miRNAs (Figure 2). The related L2 (LINE) and MIR (SINE) families, as well as DNA elements, show far more overlap with miRNA genes than is expected on the basis of their relative frequency in the genome (37 observed *vs.* 11 expected; $\chi^2 = 30.74$, $P = 3.0 \times 10^{-8}$). Most of the DNA-type elements that contribute to miRNA genes are short nonautonomous derivatives of full-length transposons known as miniature inverted-repeat transposable elements (MITEs). This includes a group of seven closely related miRNA genes (hsa-mir-548), which are all derived from the Made1 family of MITEs (PIRIYAPONGSA and JORDAN 2007). Alu (SINE) elements and LTR type TEs are generally underrepresented among TE-derived miRNA genes. Most TE-related miRNA genes are derived from a single TE insertion, but there are several examples where nested insertion events have led to the origin of a single miRNA gene from two or even three TEs (supplemental Figure 2 at <http://www.genetics.org/supplemental/>). For instance, there are two cases where a Made1 element inserted into an LTR element yielded a miRNA gene (examples 24 and 27 in supplemental Figure 2 at <http://www.genetics.org/supplemental/>), and an insertion of an Alu into a L1 (LINE) sequence also gave rise to a

miRNA gene (example 46 in supplemental Figure 2 at <http://www.genetics.org/supplemental/>).

TE-derived human miRNA genes were used as queries in BLAT searches against the human genome sequence to search for putative paralogs. There are 19 cases of TE-derived miRNA genes with closely related paralogs in the human genome (Table 2). The number of paralogs per miRNA ranges from 1, for the L1-derived hsa-mir-552, to 145, for the Made1-derived hsa-mir-548d-2.

Evolution of TE-derived miRNAs: Comparative genomic sequence data were used to assess the relative evolutionary rates of TE-derived miRNAs. This analysis was based on whole-genome sequence alignments between humans and 16 other vertebrate species. Two related approaches were used to evaluate the conservation of individual miRNA sequence sites across vertebrate genomes; the first approach results in a binary characterization of either conserved or nonconserved for each site, while the second rests on a more continuous score that relates the probability of a site being conserved. All genome sites for human miRNAs were considered using these two metrics, and the relative conservation levels for TE-derived *vs.* non-TE-derived miRNA genes were compared. A total of 32.1% of sites in TE-derived miRNAs map to the most conserved elements in the human genome. This is far greater than the ~5% of conserved sites seen for the entire human genome but significantly less than seen for non-TE-derived miRNAs, which have 63.2% conserved sites ($t = 4.39$, $P = 1.4e-5$, Student's *t*-test) (Figure 3A). When the per-site conservation probabilities of human miRNAs were measured, a similar pattern was observed. The average conservation score of TE-derived miRNAs was 0.33 compared to 0.63 for non-TE-derived miRNAs ($t = 4.37$, $P = 1.5e-5$, Student's *t*-test) (Figure 3B). In addition, the frequency distribution of the average conservation scores for all human miRNA genes reveals that, compared to non-TE-derived miRNAs, there are far more TE-derived miRNAs that show little or no conservation and fewer that are highly conserved (Figure 3C). Thus, on the whole, TE-derived miRNAs are significantly less conserved than non-TE-derived miRNAs.

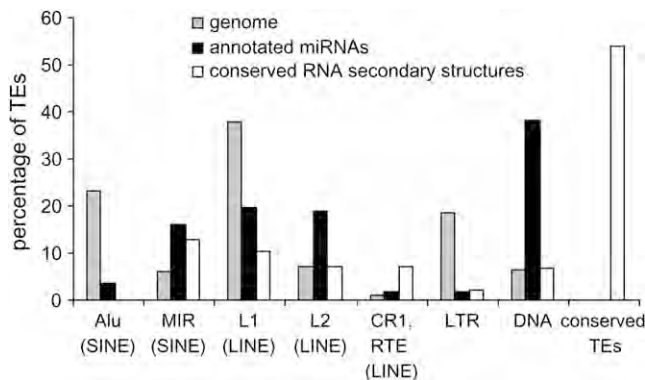


FIGURE 2.—Percentage of TE sequences among different classes and families for the human genome (shading) and for TE-derived miRNA genes (solid). Relative percentages are shown such that the total will sum to 100% for the genome and for miRNAs.

TABLE 2
Putative TE-derived miRNA paralogs

miRNA name (from miRBase)	miRBase accession no.	Colocated TE	Paralogs ^a
hsa-mir-513-1	MI0003191	MER91C	3
hsa-mir-513-2	MI0003192	MER91C	3
hsa-mir-548a-1	MI0003593	MADE1	24
hsa-mir-548a-2	MI0003598	LTR16A1, MADE1	81
hsa-mir-548a-3	MI0003612	MLT1G1, MADE1	82
hsa-mir-548b	MI0003596	MADE1	23
hsa-mir-548c	MI0003630	MADE1	124
hsa-mir-548d-1	MI0003668	MADE1	71
hsa-mir-548d-2	MI0003671	MADE1	145
hsa-mir-552	MI0003557	L1MD2	1
hsa-mir-562	MI0003568	L1MB7	2
hsa-mir-566	MI0003572	AluSg	87
hsa-mir-570	MI0003577	MADE1	48
hsa-mir-571	MI0003578	L1MA9	4
hsa-mir-579	MI0003586	MADE1, L1MB8	3
hsa-mir-603	MI0003616	MADE1	30
hsa-mir-607	MI0003620	MIR	1
hsa-mir-649	MI0003664	L1M4, MER8, AluSx	4
hsa-mir-652	MI0003667	MER91C	4

^aNumber of paralogous sequences in the human genome.

We used the frequency distribution of average conservation scores to divide TE-derived miRNAs into conserved (≥ 0.8 average conservation probability) and nonconserved (< 0.8 average conservation probability) groups. Using this criteria, there are 37 nonconserved and 18 conserved TE-derived miRNAs (Table 1). The

least-conserved TE-derived miRNAs are primate specific, having orthologous sequences in the chimpanzee only or both the chimpanzee and Rhesus genomes. Of 18 conserved miRNAs, 14 are derived from the L2 and MIR families; this is far more than would be expected on the basis of the overall frequency of L2 and MIR sequences among TE-derived miRNAs ($\chi^2 = 17.8$, $P = 3.6 \times 10^{-5}$). The conservation of L2 and MIR TE-derived miRNAs is consistent with a previous study that found many anomalously conserved L2 and MIR sequences (SILVA *et al.* 2003). Indeed, L2 and MIR are relatively ancient TE families with many sequences that inserted prior to the divergence of the human and mouse evolutionary lineages. We observed 10 of the conserved L2- and MIR-derived miRNA sequences to have orthologous sequences in the mouse genome, and there are 9 orthologous mouse miRNAs in these regions that are annotated in miRBase (Table 3). All of the 8 conserved L2 miRNAs are derived from the same region near the 3'-end of the L2 consensus sequence (approximately positions 3200–3400), while the 6 MIR-derived miRNAs are found in dispersed locations on the MIR consensus sequence.

A frequency distribution of conserved *vs.* nonconserved TE-derived miRNA genes, compared to genome-wide relative TE frequencies, reveals distinct conservation levels for miRNAs derived from particular TE classes/families (Figure 4). For instance, L2 and MIRs contribute far more conserved than nonconserved miRNAs, and the fraction of conserved L2 and MIR elements in miRNAs is much higher than seen for these same elements in the genome as a whole. DNA-type elements show the opposite pattern. There is a higher fraction of

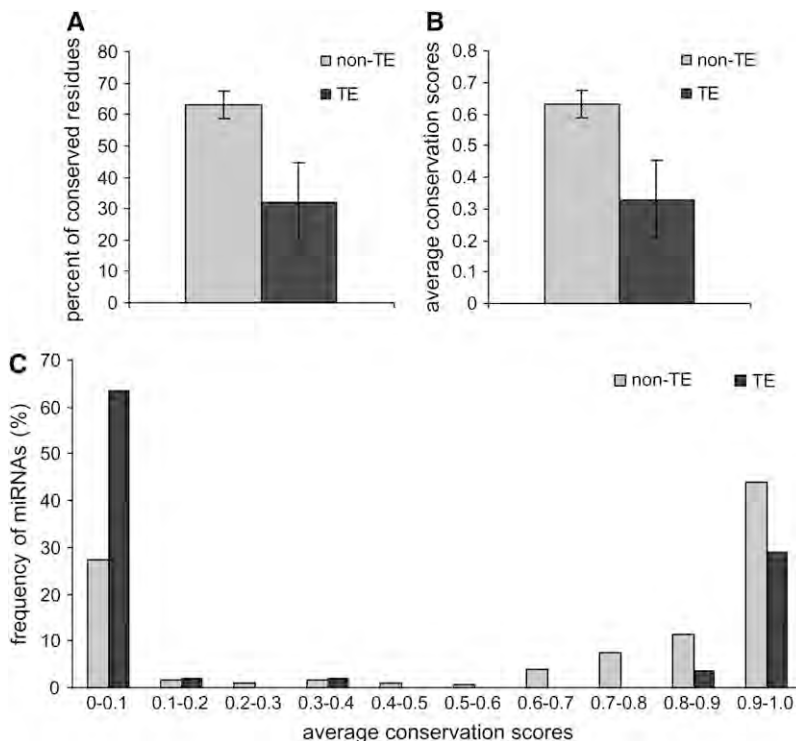


FIGURE 3.—Evolutionary conservation of human miRNA genes. (A) The percentage of conserved residues for non-TE-derived miRNAs (shading) *vs.* TE-derived miRNAs (solid) with 95% confidence intervals shown. (B) The average per-site conservation score for non-TE-derived miRNAs (shading) *vs.* TE-derived miRNAs (solid) with 95% confidence intervals shown. (C) Frequency distribution of the average per-site conservation scores for non-TE-derived miRNAs (shading) *vs.* TE-derived miRNAs (solid).

TABLE 3
Human-mouse orthologous miRNAs derived from L2 and MIR TEs

miRBase names and accession nos. for human orthologous miRNAs	Genome coordinates for human orthologous regions	Related TE sequence	miRBase names and accession nos. for mouse orthologous miRNAs	Genome coordinates for mouse orthologous regions
hsa-mir-345: MI00000825	Chromosome 14: 99843949–99844046(+)	MIR	mmu-mir-345: MI00000632	Chromosome 12: 109,284,780–109,284,874(+)
hsa-mir-130b: MI00000748	Chromosome 22: 20337593–20337674(+)	MIRm	mmu-mir-130b: MI0000408	Chromosome 16: 17,037,626–17,037,705(–)
hsa-mir-151: MI00000809	Chromosome 8: 141811845–141811934(–)	L2	mmu-mir-151: MI0000173	Gap
hsa-mir-95: MI00000097	Chromosome 4: 8057928–8058008(–)	L2	—	Gap
hsa-mir-330: MI00000803	Chromosome 19: 50834092–50834185(–)	MIRm	mmu-mir-330: MI00000607	Chromosome 7: 18,339,991–18,340,084(+)
hsa-mir-370: MI00000778	Chromosome 14: 100447229–100447303(+)	MIRm	mmu-mir-370: MI0001165	Chromosome 12: 110,066,065–110,066,139(+)
hsa-mir-325: MI00000824	Chromosome X: 76142220–76142317(–)	L2	mmu-mir-325: MI0000597	Chromosome X: 101,581,801–101,581,898(–)
hsa-mir-545: MI00003516	Chromosome X: 73423664–73423769(–)	L2	—	Chromosome X: 99,818,159–99,818,260(–)
hsa-mir-374: MI00000782	Chromosome X: 73423846–73423917(–)	L2	mmu-mir-374: MI0004125	Chromosome X: 99,818,306–99,818,361(–)
hsa-mir-28: MI00000086	Chromosome 3: 189889263–189889348(+)	L2	mmu-mir-28: MI0000690	Chromosome 16: 24,743,204–24,743,289(+)
hsa-mir-493: MI00003132	Chromosome 14: 100405150–100405238(+)	L2	—	Chromosome 12: 110,028,035–110,028,123(+)
hsa-mir-607: MI00003620	Chromosome 10: 98578416–98578511(–)	MIR	—	Gap
hsa-mir-421: MI00003685	Chromosome X: 73354937–73355021(–)	L2	—	Chromosome X: 99,775,634–99,775,718(–)
hsa-mir-378: MI00000786	Chromosome 5: 149092581–149092646(+)	MIRb	mmu-mir-378: MI0000795	Gap

“Gap” indicates no orthologous region.

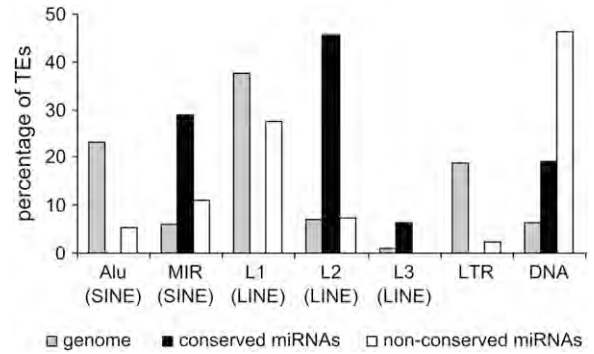


FIGURE 4.—Percentage of TE sequences among different classes and families for the human genome (shading), for conserved TE-derived miRNAs (solid), and for nonconserved TE-derived miRNAs (open). Relative percentages are shown such that the total will sum to 100% for the genome and for each group of miRNAs.

nonconserved DNA-type elements among miRNAs than is seen for the whole genome. All of the miRNAs derived from Alu and L1 elements are nonconserved.

Regulation and function: Given their high copy numbers, there is a potential for TE-derived miRNAs to regulate multiple genes via homologous target sites dispersed throughout genome. Using the miRBase target predictions, TE-derived miRNAs were found to have hundreds of putative target sites (Table 1; Figure 5A). However, while many of these target sites are also derived from TEs, in most cases the proportion of TE-derived target sites is ~10% (Table 1; Figure 5B). Thus, TE-derived miRNAs also have the potential to regulate host genes with non-TE-derived targets. The relative paucity of TE-derived target sites can be attributed, in part, to the fact that target-site prediction methods employ conservation of 3'-UTR sequences as one criteria and TEs tend to be lineage specific and nonconserved.

There are several outliers that have a substantially higher fraction of TE-derived target sites. For instance, hsa-mir-566 is derived from Alu and it has 1184 predicted targets with 948 (80%) derived from TEs. Most of these TE-derived hsa-mir-566 target sites are related to Alu insertions and this is consistent with previous studies that have found numerous putative Alu-related miRNA target sites in the human genome (DASKALOVA *et al.* 2006; SMALHEISER and TORVIK 2006).

The predicted target sites analyzed here are all putative sites and it is difficult to know with certainty whether they are actually involved in miRNA-mediated gene regulation. Another way to evaluate the regulatory potential of miRNAs is to compare the expression patterns of miRNAs to the expression patterns of the genes they are thought to regulate (FARH *et al.* 2005; STARK *et al.* 2005; HUANG *et al.* 2006; SOOD *et al.* 2006). The rationale behind the miRNA-mRNA expression pattern comparison is based on the mRNA degradation model of miRNA action. According to this model, miRNA binding to mRNA target sites causes the mRNA transcripts to be degraded.

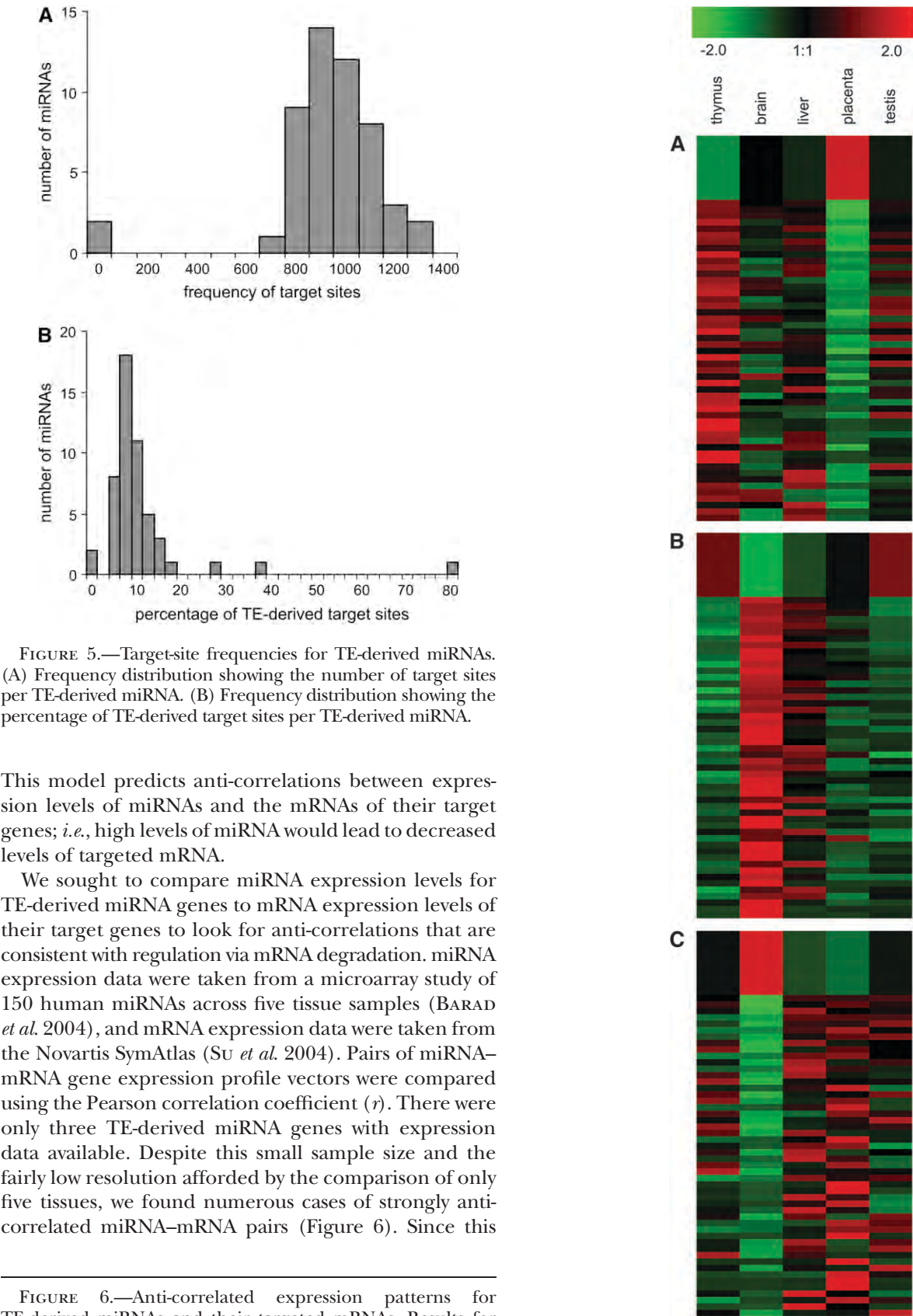


FIGURE 5.—Target-site frequencies for TE-derived miRNAs. (A) Frequency distribution showing the number of target sites per TE-derived miRNA. (B) Frequency distribution showing the percentage of TE-derived target sites per TE-derived miRNA.

This model predicts anti-correlations between expression levels of miRNAs and the mRNAs of their target genes; *i.e.*, high levels of miRNA would lead to decreased levels of targeted mRNA.

We sought to compare miRNA expression levels for TE-derived miRNA genes to mRNA expression levels of their target genes to look for anti-correlations that are consistent with regulation via mRNA degradation. miRNA expression data were taken from a microarray study of 150 human miRNAs across five tissue samples (BARAD *et al.* 2004), and mRNA expression data were taken from the Novartis SymAtlas (SU *et al.* 2004). Pairs of miRNA–mRNA gene expression profile vectors were compared using the Pearson correlation coefficient (r). There were only three TE-derived miRNA genes with expression data available. Despite this small sample size and the fairly low resolution afforded by the comparison of only five tissues, we found numerous cases of strongly anti-correlated miRNA–mRNA pairs (Figure 6). Since this

FIGURE 6.—Anti-correlated expression patterns for TE-derived miRNAs and their targeted mRNAs. Results for three TE-derived miRNAs with expression data are shown: hsa-mir-130b (A), hsa-mir-28 (B), and hsa-mir-95 (C). The top row in A–C shows the relative miRNA expression across five human tissues, and the subsequent rows show relative ex-

pression levels for targeted mRNAs. The 50 most-negative Pearson correlation coefficients (range $r = -0.99$ to -0.51 ; $P = 1.2 \times 10^{-10}$ – 1.3×10^{-1}) are shown for each plot.

anti-correlation is consistent with the mRNA degradation model of miRNA gene regulation, it provides an additional source of support for putative miRNA target sites and the regulatory action of TE-derived miRNAs.

We also evaluated the GO biological process annotations of the anti-correlated gene sets to look for overrepresented functional categories that may indicate specific functional roles for TE-derived miRNAs. The top 10% of anti-correlated mRNAs (*i.e.*, those with the lowest *r*-values) for each of the three TE-derived miRNAs with expression data were evaluated for overrepresented GO terms. The miRNA hsa-mir-130b gave the strongest signal of GO term overrepresentation; 39 of 80 genes were found to correspond to significantly overrepresented GO terms (supplemental Table 1 at <http://www.genetics.org/supplemental/>). Many of these genes correspond to metabolism and transcriptional regulation in general as well as to several negative regulators of DNA metabolism (supplemental Figure 3 at <http://www.genetics.org/supplemental/>). This negative regulation is achieved in part by chromatin remodeling, silencing, and heterochromatin formation. Thus, hsa-mir-130b may act to indirectly upregulate DNA metabolism by downregulating chromatin-based repressors.

Prediction of novel TE-derived miRNAs: The function of miRNAs, and of noncoding RNAs in general, is related to their secondary structure (MATTICK and MAKUNIN 2006). Selective constraint on such sequences often leads to compensatory mutations that maintain the base-pair interactions in the double-stranded regions of the structures, such as miRNA stem regions. Sequence alignments can be evaluated for the signal of conserved base-pair interactions as well as compensatory mutations to identify conserved, and thus presumably functionally relevant, secondary structural elements. Recent application of such techniques has led to the discovery of many novel putative regulatory RNA sequences (WASHIETL *et al.* 2005; PEDERSEN *et al.* 2006). It has even been shown that orthologous regions that are not constrained at the level of primary sequence may nevertheless encode conserved secondary structural elements (TORARINSSON *et al.* 2006). Given the contribution of TEs to experimentally characterized miRNAs shown here and elsewhere (SMALHEISER and TORVIK 2005; BORCHERT *et al.* 2006; PIRIYAPONGSA and JORDAN 2007), we sought to evaluate human TE sequences for the ability to form hairpin structures along with the signals of conserved base pairs and compensatory mutations that indicate putatively functional secondary structures. This approach provides a way to predict further contributions of TEs to miRNAs.

Human genome TE sequences were evaluated for the potential to encode conserved secondary structures (PEDERSEN *et al.* 2006) that meet the criteria of miRNA genes (BENTWICH *et al.* 2005). This approach is conservative in the sense that it relies on sequence conservation and most of the experimentally characterized TE-derived miRNAs that we observe (37 of 55) are not

evolutionarily conserved. Using this conservative approach, we found 587 human TEs with the potential to encode conserved secondary structures (supplemental Table 2 at <http://www.genetics.org/supplemental/>); 4 of these sequences corresponded to previously known human miRNAs annotated in miRBase. Evaluation of these conserved secondary structures was used to identify 85 TE-derived sequences that meet the structural criteria of putative miRNA genes, and 70 of these sequences also show evidence of being expressed (Table 4). These 70 putative TE-derived miRNA sequences meet the previously defined biogenesis, conservation, and, at least in principle, expression criteria used for the identification of miRNA genes (AMBROS *et al.* 2003).

An example of a predicted TE-derived miRNA gene is shown in Figure 7. The MER135 sequence shown is a member of a family of recently characterized nonautonomous DNA-type elements, *i.e.*, MITEs, with ~500 copies in the human genome (JURKA 2006). Since MITEs have palindromic structures with terminal inverted repeats that flank short internal regions, their expression as RNA results in the formation of the kinds of hairpins seen for pre-miRNAs. Indeed, MITEs have previously been shown to contribute miRNA genes in the Arabidopsis and human genomes (METTE *et al.* 2002; PIRIYAPONGSA and JORDAN 2007).

DISCUSSION

Abundance of TE-derived miRNAs: Noncoding regulatory RNAs, such as miRNAs, are a recently discovered class of genes, and the number of miRNA genes that exist among eukaryotic genomes is very much an open question (BEREZIKOV *et al.* 2006). Sustained efforts at high-throughput characterization of miRNA genes, based on both experimental and computational approaches, continue to result in the discovery of many novel miRNAs (BENTWICH *et al.* 2005; CUMMINS *et al.* 2006). This can be appreciated by examining the release statistics of miRBase (<ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/CURRENT/README>). Plotting the number of miRNA gene entries against the miRBase release dates suggests that the number of known miRNA genes has experienced two distinct phases of linear increase, before and after the June 2005 release, and the current rate of increase in known miRNA genes is even greater than for the initial phase (supplemental Figure 4 at <http://www.genetics.org/supplemental/>).

For the most part, the miRBase data do not include substantial numbers of computationally predicted miRNA genes. The only computational predictions represented in miRBase are highly conserved sequences that are orthologous to experimentally characterized miRNA genes in other species. Consideration of computationally identified miRNAs would suggest that miRNA gene numbers are substantially higher than currently appreciated. However, a number of computational methods for miRNA

TABLE 4
Predicted TE-derived miRNA genes

Name ^a	Coordinates ^b	Colocated TE	Expression data ^c
3715_0_+_61	Chromosome 1: 3131597–3131629(+)	MER121	EST/mRNA/KG/RS
15086_0_–_78	Chromosome 1: 15041842–15041859(–)	HAL1	EST/mRNA/KG/RS
25288_0_–_83	Chromosome 1: 23621848–23621877(–)	MIRb	EST/mRNA/KG/RS
30647_0_+_38	Chromosome 1: 27752374–27752433(+)	MIRb	EST/mRNA/KG/RS
52664_0_–_50	Chromosome 1: 44571346–44571464(–)	Eulor9A	EST/mRNA/KG/RS
67626_0_–_76	Chromosome 1: 57127400–57127465(–)	Eulor1	EST/mRNA/KG/RS
85615_0_+_83	Chromosome 1: 76474930–76474947(+)	MIRb	EST/mRNA/KG/RS
120809_0_+_79	Chromosome 1: 111021701–111021719(+)	MIR	EST/mRNA
122080_0_–_62	Chromosome 1: 112177611–112177631(–)	MIR	EST/mRNA/KG/RS
124780_0_–_66	Chromosome 1: 114214379–114214407(–)	MIRb	EST/mRNA/KG/RS
154818_0_–_64	Chromosome 1: 162825371–162825437(–)	MER135	EST/mRNA/KG/RS
188052_1_–_92	Chromosome 1: 198460508–198460590(–)	Eulor3	—
204532_0_–_104	Chromosome 1: 211522027–211522054(–)	UCON31	EST
230542_0_–_67	Chromosome 1: 244286075–244286098(–)	L1MB3	EST/mRNA/KG/RS
1231553_0_+_75	Chromosome 2: 67238894–67239028(+)	Eulor4	EST/mRNA
1258257_0_+_85	Chromosome 2: 104314401–104314489(+)	MER134	—
1361323_0_+_57	Chromosome 2: 213067475–213067509(+)	Eulor5A	EST/mRNA/KG/RS
1573547_0_+_44	Chromosome 3: 61643441–61643518(+)	MER126	EST/mRNA/KG/RS
1573643_0_+_95	Chromosome 3: 61718341–61718381(+)	MER134	EST/mRNA/KG/RS
1620066_0_–_64	Chromosome 3: 116298434–116298458(–)	Eulor1	EST/mRNA/KG/RS
1651767_0_+_52	Chromosome 3: 146074810–146074873(+)	Eulor3	—
1668216_0_–_58	Chromosome 3: 168436231–168436447(–)	MER126	—
1730972_0_–_56	Chromosome 4: 46681709–46681733(–)	L1ME3B	EST/mRNA/KG/RS
1747758_0_–_63	Chromosome 4: 74275595–74275629(–)	L1M5	EST/mRNA/KG/RS
1757379_0_+_70	Chromosome 4: 85466757–85466855(+)	MER134	—
1827751_0_+_75	Chromosome 4: 181988895–181988914(+)	MIRb	EST
1830405_0_+_49	Chromosome 4: 183690755–183690850(+)	MER135	EST/mRNA/RS
1873731_0_+_53	Chromosome 5: 58495675–58495729(+)	UCON9	EST/mRNA/KG/RS
1902777_0_+_53	Chromosome 5: 90643387–90643420(+)	AmnSINE1_GG	EST/mRNA
1920501_0_+_72	Chromosome 5: 113735156–113735173(+)	L2	EST/mRNA/KG/RS
1966281_0_+_83	Chromosome 5: 156681824–156681841(+)	MIR3	EST/mRNA/KG/RS
1975838_0_–_80	Chromosome 5: 165688874–165688944(–)	Eulor5A	—
1979031_0_+_61	Chromosome 5: 167506770–167506888(+)	Eulor9A	EST/mRNA/RS
1987527_0_+_59	Chromosome 5: 175727565–175727628(+)	L2	EST/mRNA/KG/RS
2000476_0_–_85	Chromosome 6: 8499794–8499914(–)	Eulor6C	EST/mRNA
2031067_0_+_44	Chromosome 6: 39048083–39048162(+)	Eulor5A	EST/mRNA/KG/RS
2075048_0_–_91	Chromosome 6: 94484941–94484963(–)	ERVLE	EST/mRNA
2115069.5_0_+_82	Chromosome 6: 141179709–141179763(+)	Eulor5B	—
2165103_0_+_104	Chromosome 7: 28447122–28447144(+)	MER121	EST/mRNA/KG/RS
2195049_0_+_117	Chromosome 7: 73161289–73161306(+)	MIR3	EST/mRNA/KG/RS
2232211_0_+_45	Chromosome 7: 113190696–113190791(+)	Eulor6B	—
2247695_1_+_65	Chromosome 7: 129521966–129521985(+)	L1ME4a	EST/mRNA/KG/RS
2265159_0_+_85	Chromosome 7: 146833245–146833271(+)	UCON4	EST/mRNA/KG/RS
2330918_0_–_108	Chromosome 8: 79081399–79081462(–)	Eulor3	—
2344217_0_+_65	Chromosome 8: 97188471–97188580(+)	MER135	EST
2348773_0_+_51	Chromosome 8: 102229956–102230022(+)	Charlie9	—
2401146_0_–_96	Chromosome 9: 16787222–16787246(–)	MIR	EST/mRNA/KG/RS
2421368_0_–_79	Chromosome 9: 37811135–37811158(–)	L1MC4a	EST/mRNA/KG/RS
2426661_0_+_64	Chromosome 9: 70297285–70297306(+)	MER91A	EST/KG/RS
2455634_0_–_64	Chromosome 9: 105918396–105918420(–)	MER5A	EST/mRNA/KG/RS
2469999_0_+_79	Chromosome 9: 118715772–118715795(+)	UCON11	EST/mRNA/KG/RS
2500550_0_–_83	Chromosome X: 10899595–10899617(–)	L4	EST/mRNA/KG/RS
2519737_0_+_67	Chromosome X: 24557155–24557175(+)	L1ME4a	EST/mRNA/KG/RS
2598753_0_+_171	Chromosome X: 123865376–123865447(+)	Eulor11	EST/mRNA/KG/RS
2607024_0_–_68	Chromosome X: 131689852–131689873(–)	L1MB5	EST/mRNA/KG/RS
2625375_0_+_86	Chromosome X: 152562536–152562556(+)	L2	EST/mRNA/KG/RS
276291_0_+_66	Chromosome 10: 62836157–62836220(+)	L1M5	—
285555_0_+_63	Chromosome 10: 72980870–72980944(+)	MER125	EST/mRNA/KG/RS

(continued)

TABLE 4
(Continued)

Name ^a	Coordinates ^b	Colocated TE	Expression data ^c
334961_0_+_78	Chromosome 10: 117579937–117579954(+)	L2	EST/mRNA/KG/RS
335779_0_+_54	Chromosome 10: 118027456–118027512(+)	Eulor6D	EST
377681_0_+_96	Chromosome 11: 19331037–19331062(+)	L3	mRNA/KG
425555_0_+_71	Chromosome 11: 71985685–71985701(+)	MIR	EST/mRNA/KG/RS
438439_0_+_83	Chromosome 11: 83316376–83316398(+)	L2	EST/mRNA/KG/RS
486187_0_+_68	Chromosome 11: 130861130–130861151(+)	MIRb	EST/mRNA/KG/RS
487071_2_+_103	Chromosome 11: 131453921–131453949(+)	MER122	EST/mRNA/KG/RS
492576_0_–_95	Chromosome 12: 2125422–2125443(–)	MIRb	mRNA/KG/RS
533638.0_0_–_122	Chromosome 12: 50492331–50492353(–)	MIRb	mRNA/KG
542148_0_–_83	Chromosome 12: 55246557–55246574(–)	LTR37B	EST/mRNA/KG/RS
551096_0_–_85	Chromosome 12: 64538090–64538148(–)	Eulor5A	EST/mRNA/KG/RS
596947_0_+_93	Chromosome 12: 115505370–115505426(+)	MER123	EST/mRNA
697653_0_+_69	Chromosome 14: 33093444–33093479(+)	UCON11	EST/mRNA/KG/RS
700890_0_–_65	Chromosome 14: 35855217–35855366(–)	Eulor6A	EST/mRNA/KG/RS
775713_0_+_77	Chromosome 15: 25703141–25703162(+)	L1MCc	EST/mRNA/KG/RS
787092_0_–_65	Chromosome 15: 35993736–35993832(–)	Eulor5A	—
896537_0_+_81	Chromosome 16: 30749660–30749680(+)	MIR	EST
928869_0_+_74	Chromosome 16: 70304015–70304037(+)	MIR3	EST/mRNA/KG/RS
976169_0_+_86	Chromosome 17: 24040248–24040268(+)	L1ME4a	EST/mRNA/KG/RS
989909_0_+_100	Chromosome 17: 34009010–34009024(+)	MIR3	EST/mRNA/KG/RS
1000039.8_0_+_109	Chromosome 17: 39468501–39468532(+)	L1MC4	EST/mRNA/KG/RS
1077028_0_–_58	Chromosome 18: 33875730–33875789(–)	MIRb	—
1105916_0_–_78	Chromosome 18: 71369451–71369514(–)	UCON11	—
1435354_0_–_79	Chromosome 20: 44235903–44235921(–)	MIR	EST/mRNA/KG/RS
1443968_0_–_61	Chromosome 20: 53838763–53838824(–)	UCON29	—
1466070_0_–_70	Chromosome 21: 33853177–33853203(–)	L2	EST/mRNA
1496941_0_+_79	Chromosome 22: 35289947–35289989(+)	L1MC4	EST

^a Name of the EvoFold locus from the hg18 UCSC Genome Browser annotation. The last field in the name corresponds to the EvoFold score.

^b Genome coordinates and strand of the EvoFold locus.

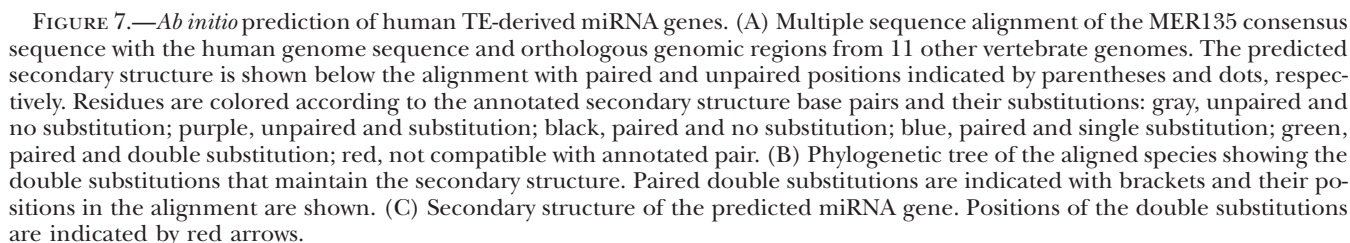
^c Source of the expression data for the locus: KG, UCSC Genome Browser known gene annotation; RS, NCBI RefSeq annotation.

prediction do not consider TE-derived miRNAs (BENTWICH *et al.* 2005; LINDOW and KROGH 2005; NAM *et al.* 2005; LI *et al.* 2006). This is because, mainly for reasons of tractability, one of the first steps in computational analysis of eukaryotic genome sequences is the exclusion of repetitive DNA by RepeatMasking. TEs will also tend to be excluded from predictions based solely on conservation between species because they are rapidly evolving and lineage-specific genomic elements. This is underscored by the fact that the set of TE-derived human miRNAs that we identify here is enriched for genes experimentally characterized in humans (93% for TE-derived *vs.* 81% for non-TE-derived miRNAs; $\chi^2 = 4.76$, $P = 0.03$).

The factors described above that suggest the exclusion of TE-derived miRNAs led us to speculate as to how many more miRNA genes would be discovered if TE sequences were not eliminated from consideration *a priori*. To investigate this, we employed our own *ab initio* computational approach to try and predict TE-derived miRNA sequences. Application of this method to the human genome revealed 587 cases of human TE sequences that encode conserved RNA secondary structures, 85 of which are most likely to represent *bona fide*

miRNA genes. Fifteen of the TE-derived miRNA genes that we predicted using this approach overlap with previous miRNA computational predictions (BEREZIKOV *et al.* 2005; PEDERSEN *et al.* 2006) as well as experimentally characterized miRNAs from miRBase.

Conservation of TE-derived miRNAs: Many miRNA genes are evolutionarily conserved and may have functional orthologs in multiple species. Indeed, sequence conservation is one of the criteria used to aid the computational discovery of miRNAs. While the TE-derived miRNA genes analyzed here are less conserved, on average, than non-TE-derived miRNAs, there are a number of well-conserved miRNAs that evolved from TE sequences (Table 1). The majority of these conserved miRNAs are related to the ancient L2 and MIR TE families, and some of these sequences have been previously identified (SMALHEISER and TORVIK 2005). This is particularly interesting because numerous L2 and MIR sequences have been shown to be anomalously conserved between the human and mouse genomes (SILVA *et al.* 2003). Specifically, SILVA *et al.* (2003) demonstrated that many L2 and MIR sequences found in orthologous human–mouse intergenic regions were present in the common ancestor of the two species and,



As in the case of L2 and MIR (SILVA *et al.* 2003), comparative genomic approaches are used to infer functionally important genomic regions, particularly noncoding regions, by virtue of their high sequence conservation

(ZHANG and GERSTEIN 2003). It is becoming increasingly apparent that a number of such highly conserved genomic sequences correspond to TEs (BEJERANO *et al.* 2006; KAMAL *et al.* 2006; NISHIHARA *et al.* 2006; XIE *et al.* 2006). While enhancer activity has been demonstrated for one of these conserved TEs (BEJERANO *et al.* 2006), for the most part, the specific function encoded by conserved TE sequences remains unknown. The collection of conserved TE sequences recently assembled by Repbase corresponds to <1% of all human genome TEs, but these sequences contribute >50% of all TE-encoded conserved secondary structures that we detected (Figure 2). Thus, our results

suggest that many conserved TE sequences may encode miRNAs or perhaps other noncoding regulatory or structural RNAs.

Lineage-specific effects of TE-derived miRNAs: Most of the TE-derived miRNAs analyzed here are not evolutionarily conserved (Table 1). This is not surprising when you consider that TEs are the most lineage-specific and nonconserved elements found in eukaryotic genomes (LANDER *et al.* 2001). The overrepresentation of non-conserved sequences among TE-derived miRNAs is also consistent with previous work that has shown TE-derived *cis*-regulatory binding sites to be more divergent than non-TE-derived *cis* sites (MARIÑO-RAMIREZ *et al.* 2005). From a practical perspective, this means that computational discovery methods that employ conservation as a criterion will necessarily overlook many TE-derived regulatory sequences. In terms of evolution, this means that the greatest differences between eukaryotic genomes will correspond to TE sequences. In this sense, TEs can be considered as drivers of genome diversification. This may be uninteresting if TEs serve only to replicate themselves and do not play any role for their host genomes as the selfish DNA theory of TEs holds (DOOLITTLE and SAPIENZA 1980; ORGEL and CRICK 1980). However, if some TEs are in fact functionally relevant to their hosts, as we have shown here for the case of TE-derived miRNAs, then their divergence may have important evolutionary implications. Indeed, TE-derived regulatory sequences may be particularly prone to contribute to regulatory differences among species that lead to lineage-specific phenotypes. This has been shown for the case of TE-derived regulatory sequences that are associated with high levels of expression divergence between humans and mice (MARIÑO-RAMIREZ and JORDAN 2006).

While most computational efforts to discover non-coding regulatory sequences have focused on conserved genomic elements, recent studies have begun to emphasize rapidly evolving regions as well (POLLARD *et al.* 2006a,b; PRABHAKAR *et al.* 2006). The rationale behind this is the notion that rapidly evolving regulatory regions may yield species-specific differences. An emphasis on the discovery of TE-derived regulatory sequences would complement current approaches to the discovery of rapidly evolving regulatory regions that are likely to contribute to the phenotypic divergence among species.

Genome defense and global gene regulatory mechanisms: Finally, we speculate that our results point to a connection between genome defense mechanisms necessitated by TEs and the emergence of global gene regulatory systems that may have allowed for the complex regulatory phenotypes characteristic of multicellular eukaryotes. TE insertions are highly deleterious and, as a consequence, a number of global gene-silencing mechanisms, including methylation (YODER *et al.* 1997), imprinting (MCDONALD *et al.* 2005), and heterochromatin (LIPPMAN *et al.* 2004), may have evolved originally as TE defense mechanisms. siRNAs are also thought

to have evolved as a defense mechanism against TEs (MATZKE *et al.* 2000; VASTENHOUW and PLASTERK 2004; SLOTKIN *et al.* 2005), and the results reported here and elsewhere (SMALHEISER and TORVIK 2005; BORCHERT *et al.* 2006; PIRIYAPONGSA and JORDAN 2007) indicate that miRNAs can emerge from TEs as well. More recently, an analogous TE defense mechanism based on small RNAs complementary to TEs in *Drosophila* has been reported (BRENNER *et al.* 2007). Apparently, different RNA interference systems may have evolved convergently on multiple occasions to help silence TEs. Later, these regulatory mechanisms could have been co-opted to exert controlling effects over thousands of host genes as is the case for miRNAs. The evolution of such complex gene regulatory systems can be considered non-adaptive (LYNCH 2007) in the sense that they did not evolve by virtue of selection for the role that they play now. However, neither did these global regulatory mechanisms evolve passively since they were swept to fixation by selective pressure to defend against TEs. Therefore, the emergence of TE-related global regulatory systems, exemplified by RNA interference, can be considered to be exaptations (GOULD and VRBA 1982) driven by the internal mutational dynamics (STOLTZFUS 2006) of the genome.

The authors thank Nalini Polavarapu and Ahsan Huda for technical support and helpful comments. Jittima Piriyaongsa is supported by the Ministry of Science and Technology of Thailand. I. King Jordan is supported by the School of Biology at the Georgia Institute of Technology. This research was supported in part by the intramural research program of the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health.

LITERATURE CITED

- AMBROS, V., 2004 The functions of animal microRNAs. *Nature* **431**: 350–355.
- AMBROS, V., B. BARTEL, D. P. BARTEL, C. B. BURGE, J. C. CARRINGTON *et al.*, 2003 A uniform system for microRNA annotation. *RNA* **9**: 277–279.
- ASHBURNER, M., C. A. BALL, J. A. BLAKE, D. BOTSTEIN, H. BUTLER *et al.*, 2000 Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat. Genet.* **25**: 25–29.
- BARAD, O., E. MEIRI, A. AVNIEL, R. AHARONOV, A. BARZILAI *et al.*, 2004 MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues. *Genome Res.* **14**: 2486–2494.
- BARTEL, D. P., 2004 MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**: 281–297.
- BEJERANO, G., C. B. LOWE, N. AHITUV, B. KING, A. SIEPEL *et al.*, 2006 A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**: 87–90.
- BENTWICH, I., A. AVNIEL, Y. KAROV, R. AHARONOV, S. GILAD *et al.*, 2005 Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* **37**: 766–770.
- BEREZIKOV, E., V. GURYEV, J. VAN DE BELT, E. WIENHOLDS, R. H. PLASTERK *et al.*, 2005 Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**: 21–24.
- BEREZIKOV, E., E. CUPPEN and R. H. PLASTERK, 2006 Approaches to microRNA discovery. *Nat. Genet.* **38**(Suppl.): S2–S7.
- BLANCHETTE, M., W. J. KENT, C. RIEMER, L. ELNITSKI, A. F. SMIT *et al.*, 2004 Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**: 708–715.

- BORCHERT, G. M., W. LANIER and B. L. DAVIDSON, 2006 RNA polymerase III transcribes human microRNAs. *Nat. Struct. Mol. Biol.* **13**: 1097–1101.
- BRENNECKE, J., D. R. HIPFNER, A. STARK, R. B. RUSSELL and S. M. COHEN, 2003 bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene *hid* in *Drosophila*. *Cell* **113**: 25–36.
- BRENNECKE, J., A. A. ARAVIN, A. STARK, M. DUS, M. KELLIS *et al.*, 2007 Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**: 1089–1103.
- BRITTON, R. J., 1996 DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci. USA* **93**: 9374–9377.
- CHEN, C. Z., L. LI, H. F. LODISH and D. P. BARTEL, 2004 MicroRNAs modulate hematopoietic lineage differentiation. *Science* **303**: 83–86.
- CUMMINS, J. M., Y. HE, R. J. LEARY, R. PAGLIARINI, L. A. DIAZ, JR. *et al.*, 2006 The colorectal microRNAome. *Proc. Natl. Acad. Sci. USA* **103**: 3687–3692.
- DASKALOVA, E., V. BAEV, V. RUSINOV and I. MINKOV, 2006 3' UTR-located Alu elements: donors of potential miRNA target sites and mediators of network miRNA-based regulatory interactions. *Evol. Bioinform. Online* **2**: 99–116.
- DOOLITTLE, W. F., and C. SAPIENZA, 1980 Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**: 601–603.
- ENRIGHT, A. J., B. JOHN, U. GAUL, T. TUSCHL, C. SANDER *et al.*, 2003 MicroRNA targets in *Drosophila*. *Genome Biol.* **5**: R1.
- FARH, K. K., A. GRIMSON, C. JAN, B. P. LEWIS, W. K. JOHNSTON *et al.*, 2005 The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310**: 1817–1821.
- GOULD, S. J., and E. S. VRBA, 1982 Exaptation: a missing term in the science of form. *Paleobiology* **8**: 4–15.
- GRIFFITHS-JONES, S., R. J. GROCOCK, S. VAN DONGEN, A. BATEMAN and A. J. ENRIGHT, 2006 miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**: D140–D144.
- HOFACKER, I. L., W. FONTANA, P. F. STADLER, S. BONHOEFFER, M. TACKER *et al.*, 1994 Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**: 167–188.
- HUANG, J. C., Q. D. MORRIS and B. J. FREY, 2006 Detecting microRNA targets by linking sequence, microRNA and gene expression data, pp. 114–129 in *RECOMB 2006*, edited by A. APOSTOLICO, C. GUERRA, S. ISTRAIL, P. A. PEVZNER and M. S. WATERMAN. Springer-Verlag, Venice, Italy.
- JORDAN, I. K., I. B. ROGOZIN, G. V. GLAZKO and E. V. KOONIN, 2003 Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* **19**: 68–72.
- JURKA, J., 2006 MER135: conserved mammalian repeat, probably derived from a non-autonomous DNA transposon. *Rebase Rep.* **6**: 388.
- JURKA, J., V. V. KAPITONOV, A. PAVLICEK, P. KLONOWSKI, O. KOHANY *et al.*, 2005 Rebase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**: 462–467.
- KAMAL, M., X. XIE and E. S. LANDER, 2006 A large family of ancient repeat elements in the human genome is under strong selection. *Proc. Natl. Acad. Sci. USA* **103**: 2740–2745.
- KAROLCHIK, D., A. S. HINRICHS, T. S. FUREY, K. M. ROSKIN, C. W. SUGNET *et al.*, 2004 The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**: D493–D496.
- KENT, W. J., 2002 BLAT: the BLAST-like alignment tool. *Genome Res.* **12**: 656–664.
- KENT, W. J., C. W. SUGNET, T. S. FUREY, K. M. ROSKIN, T. H. PRINGLE *et al.*, 2002 The human genome browser at UCSC. *Genome Res.* **12**: 996–1006.
- KENT, W. J., R. BAERTSCH, A. HINRICHS, W. MILLER and D. HAUSSLER, 2003 Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* **100**: 11484–11489.
- KIDWELL, M. G., and D. R. LISCH, 2001 Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int. J. Org. Evolution* **55**: 1–24.
- LAGOS-QUINTANA, M., R. RAUHUT, W. LENDECKEL and T. TUSCHL, 2001 Identification of novel genes coding for small expressed RNAs. *Science* **294**: 853–858.
- LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- LAU, N. C., L. P. LIM, E. G. WEINSTEIN and D. P. BARTEL, 2001 An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**: 858–862.
- LEE, R. C., and V. AMBROS, 2001 An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294**: 862–864.
- LEE, R. C., R. L. FEINBAUM and V. AMBROS, 1993 The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**: 843–854.
- LI, S. C., C. Y. PAN and W. C. LIN, 2006 Bioinformatic discovery of microRNA precursors from human ESTs and introns. *BMC Genomics* **7**: 164.
- LINDOW, M., and A. KROGH, 2005 Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics* **6**: 119.
- LIPPMAN, Z., A. V. GENDREL, M. BLACK, M. W. VAUGHN, N. DEDHIA *et al.*, 2004 Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**: 471–476.
- LYNCH, M., 2007 *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA.
- MARIÑO-RAMÍREZ, L., and I. K. JORDAN, 2006 Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol. Direct* **1**: 20.
- MARIÑO-RAMÍREZ, L., K. C. LEWIS, D. LANDSMAN and I. K. JORDAN, 2005 Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet. Genome Res.* **110**: 333–341.
- MATTICK, J. S., and I. V. MAKUNIN, 2006 Non-coding RNA. *Hum. Mol. Genet.* **15** Spec. No. 1: R17–R29.
- MATZKE, M. A., M. F. METTE and A. J. MATZKE, 2000 Transgene silencing by the host genome defense: implications for the evolution of epigenetic control mechanisms in plants and vertebrates. *Plant Mol. Biol.* **43**: 401–415.
- MCDONALD, J. F., M. A. MATZKE and A. J. MATZKE, 2005 Host defenses to transposable elements and the evolution of genomic imprinting. *Cytogenet. Genome Res.* **110**: 242–249.
- METTE, M. F., J. VAN DER WINDEN, M. MATZKE and A. J. MATZKE, 2002 Short RNAs can identify new candidate transposable element families in *Arabidopsis*. *Plant Physiol.* **130**: 6–9.
- MILLER, W. J., S. HAGEMANN, E. REITER and W. PINSKER, 1992 P-element homologous sequences are tandemly repeated in the genome of *Drosophila guanche*. *Proc. Natl. Acad. Sci. USA* **89**: 4018–4022.
- NAM, J. W., K. R. SHIN, J. HAN, Y. LEE, V. N. KIM *et al.*, 2005 Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.* **33**: 3570–3581.
- NISHIHARA, H., A. F. SMIT and N. OKADA, 2006 Functional noncoding sequences derived from SINES in the mammalian genome. *Genome Res.* **16**: 864–874.
- ORGE, L. E., and F. H. CRICK, 1980 Selfish DNA: the ultimate parasite. *Nature* **284**: 604–607.
- PASQUINELLI, A. E., B. J. REINHART, F. SLACK, M. Q. MARTINDALE, M. I. KURODA *et al.*, 2000 Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. *Nature* **408**: 86–89.
- PEDERSEN, J. S., G. BEJERANO, A. SIEPEL, K. ROSENBLUM, K. LINDBLAD-TOH *et al.*, 2006 Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2**: e33.
- PIRIYAONGSA, J., and I. K. JORDAN, 2007 A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE* **2**: e203.
- POLLARD, K. S., S. R. SALAMA, B. KING, A. D. KERN, T. DRESZER *et al.*, 2006a Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2**: e168.
- POLLARD, K. S., S. R. SALAMA, N. LAMBERT, M. A. LAMBOT, S. COPPENS *et al.*, 2006b An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**: 167–172.
- PRABHAKAR, S., J. P. NOONAN, S. PAABO and E. M. RUBIN, 2006 Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**: 786.
- REINHART, B. J., F. J. SLACK, M. BASSON, A. E. PASQUINELLI, J. C. BETTINGER *et al.*, 2000 The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**: 901–906.
- SIEPEL, A., G. BEJERANO, J. S. PEDERSEN, A. S. HINRICHS, M. HOU *et al.*, 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**: 1034–1050.

- SILVA, J. C., S. A. SHABALINA, D. G. HARRIS, J. L. SPOUGE and A. S. KONDRASHOVI, 2003 Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet. Res.* **82**: 1–18.
- SLOTKIN, R. K., M. FREELING and D. LISCH, 2005 Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. *Nat. Genet.* **37**: 641–644.
- SMALHEISER, N. R., and V. I. TORVIK, 2005 Mammalian microRNAs derived from genomic repeats. *Trends Genet.* **21**: 322–326.
- SMALHEISER, N. R., and V. I. TORVIK, 2006 Alu elements within human mRNAs are probable microRNA targets. *Trends Genet.* **22**: 532–536.
- SMIT, A. F. A., R. HUBLEY and P. GREEN, 1996–2004 RepeatMasker Open-3.0 (<http://www.repeatmasker.org>).
- SOOD, P., A. KREK, M. ZAVOLAN, G. MACINO and N. RAJEWSKY, 2006 Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc. Natl. Acad. Sci. USA* **103**: 2746–2751.
- STARK, A., J. BRENNKE, N. BUSHATI, R. B. RUSSELL and S. M. COHEN, 2005 Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell* **123**: 1133–1146.
- STOLTZFUS, A., 2006 Mutationism and the dual causation of evolutionary change. *Evol. Dev.* **8**: 304–317.
- STURN, A., J. QUACKENBUSH and Z. TRAJANOSKI, 2002 Genesis: cluster analysis of microarray data. *Bioinformatics* **18**: 207–208.
- SU, A. I., T. WILTSHIRE, S. BATALOV, H. LAPP, K. A. CHING *et al.*, 2004 A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**: 6062–6067.
- TORARINSSON, E., M. SAWERA, J. H. HAVGAARD, M. FREDHOLM and J. GORODKIN, 2006 Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.* **16**: 885–889.
- VAN DE LAGEMAAT, L. N., J. R. LANDRY, D. L. MAGER and P. MEDSTRAND, 2003 Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* **19**: 530–536.
- VASTENHOUW, N. L., and R. H. PLASTERK, 2004 RNAi protects the *Caenorhabditis elegans* germline against transposition. *Trends Genet.* **20**: 314–319.
- VOLFF, J. N., 2006 Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays* **28**: 913–922.
- WASHIETI, S., I. L. HOFACKER and P. F. STADLER, 2005 Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* **102**: 2454–2459.
- XIE, X., M. KAMAL and E. S. LANDER, 2006 A family of conserved noncoding elements derived from an ancient transposable element. *Proc. Natl. Acad. Sci. USA* **103**: 11659–11664.
- YODER, J. A., C. P. WALSH and T. H. BESTOR, 1997 Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**: 335–340.
- ZHANG, B., D. SCHMOYER, S. KIROV and J. SNODDY, 2004 GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* **5**: 16.
- ZHANG, Z., and M. GERSTEIN, 2003 Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J. Biol.* **2**: 11.

Communicating editor: D. VOYTAS

ORIGINAL ARTICLE

TLX1/HOX11-induced hematopoietic differentiation blockade

I Riz¹, SS Akimov¹, SS Eaker², KK Baxter^{1,3}, HJ Lee^{1,4}, L Mariño-Ramírez⁵, D Landsman⁵, TS Hawley⁶ and RG Hawley^{1,3}

¹Department of Anatomy and Cell Biology, The George Washington University Medical Center, Washington, DC, USA;

²NanoDetection Technology, Knoxville, TN, USA; ³Molecular Medicine Program, The George Washington University Medical Center, Washington, DC, USA; ⁴Genomics and Bioinformatics Program, The George Washington University Medical Center, Washington, DC, USA; ⁵Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, USA and ⁶Flow Cytometry Core Facility, The George Washington University Medical Center, Washington, DC, USA

Aberrant expression of the human homeobox-containing proto-oncogene *TLX1/HOX11* inhibits hematopoietic differentiation programs in a number of murine model systems. Here, we report the establishment of a murine erythroid progenitor cell line, iEBHX1S-4, developmentally arrested by regulatable *TLX1* expression. Extinction of *TLX1* expression released the iEBHX1S-4 differentiation block, allowing erythropoietin-dependent acquisition of erythroid markers and hemoglobin synthesis. Coordinated activation of erythroid transcriptional networks integrated by the acetyltransferase co-activator CREB-binding protein (CBP) was suggested by bioinformatic analysis of the upstream regulatory regions of several conditionally induced iEBHX1S-4 gene sets. In accord with this notion, CBP-associated acetylation of GATA-1, an essential regulator of erythroid differentiation, increased concomitantly with *TLX1* downregulation. Coimmunoprecipitation experiments and glutathione-S-transferase pull-down assays revealed that *TLX1* directly binds to CBP, and confocal laser microscopy demonstrated that the two proteins partially colocalize at intranuclear sites in iEBHX1S-4 cells. Notably, the distribution of CBP in conditionally blocked iEBHX1S-4 cells partially overlapped with chromatin marked by a repressive histone methylation pattern, and downregulation of *TLX1* coincided with exit of CBP from these heterochromatic regions. Thus, we propose that *TLX1*-mediated differentiation arrest may be achieved in part through a mechanism that involves redirection of CBP and/or its sequestration in repressive chromatin domains.

Oncogene (2007) 26, 4115–4123; doi:10.1038/sj.onc.1210185; published online 8 January 2007

Keywords: *TLX1/HOX11* oncogene; erythropoiesis; conditional differentiation block; CBP; GATA-1; repressive chromatin domains

Introduction

The murine ortholog of human *TLX1* (previously known as *HOX11* and *TCL3*), which is a member of the dispersed NK homeobox gene family, is essential for splenogenesis and the proper development of certain sensory neurons. Although *TLX1* is not expressed in the hematopoietic system, its inappropriate activation – frequently owing to translocations involving T-cell receptor (TCR) gene loci – is a recurrent event in human T-cell acute lymphoblastic leukemia (T-ALL) (Owens and Hawley, 2002). We previously reported that enforced expression of *TLX1* immortalizes various myeloerythroid progenitors in murine bone marrow, yolk sac and embryonic stem cell (ESC)-derived embryoid bodies (Hawley *et al.*, 1994, 1997; Keller *et al.*, 1998; Owens *et al.*, 2003). Based on these results, we postulated that *TLX1* exerts its T-cell oncogenic effects in part by impeding hematopoietic differentiation programs. In support of this hypothesis, we recently demonstrated that retroviral expression of *TLX1* disrupted T-cell-directed differentiation of primary murine fetal liver precursors and human cord blood CD34⁺ stem/progenitor cells in fetal thymic organ cultures (Owens *et al.*, 2006).

The mechanism of the *TLX1*-mediated differentiation block and, by extension, the manner in which deregulated *TLX1* expression induces neoplastic conversion remain to be elucidated (Hawley *et al.*, 1997). Several lines of evidence indicate that *TLX1* functions as a transcriptional regulator that can either activate or repress gene expression via direct or indirect modes of action (Dear *et al.*, 1993; Greene *et al.*, 1998; Owens *et al.*, 2003; Riz and Hawley, 2005). A plausible assumption has been that some *TLX1* transcriptional activity is mediated by selective recognition of DNA sequences (Dear *et al.*, 1993; Allen *et al.*, 2000). Of note, however, although several genes downstream of *TLX1* transcriptional cascades have been identified to date, in no instance has direct binding of *TLX1* to the promoter sequences of primary target genes been demonstrated. On the contrary, *TLX1* has been shown in many instances to indirectly regulate gene expression *in vivo* through cooperative protein–protein interactions with other molecules (Kawabe *et al.*, 1997; Zhang *et al.*, 1999; Riz and Hawley, 2005).

Correspondence: Dr RG Hawley, Department of Anatomy and Cell Biology, The George Washington University Medical Center, 2300 I Street NW, Washington, DC 20037, USA.

E-mail: rghawley@gwu.edu

Received 7 February 2006; revised 23 October 2006; accepted 23 October 2006; published online 8 January 2007

Recent investigations have identified new recurrent TCR chromosomal translocations in human T-ALL that deregulate the *HOXA* cluster of the HOX homeobox gene family (Soulier *et al.*, 2005). Genome-wide expression analysis showed that the *HOXA*-translocated cases shared multiple transcriptional networks with *TLX1*⁺ T-ALL samples (Soulier *et al.*, 2005), suggesting a common mechanism underlying these malignancies. Many HOX proteins have been reported to interact with the ubiquitously expressed acetyltransferase co-activator CREB-binding protein (CBP) and its paralog p300 (Shen *et al.*, 2001). In particular, all 14 HOX proteins tested in one study, representing 11 of the 13 paralogous groups, were shown to associate with CBP in a DNA-binding-independent manner and inhibit CBP acetyltransferase activity (Shen *et al.*, 2001). CBP regulates gene expression in most if not all cell types, functioning as a molecular integrator linking a large number of transcription factors to the basal transcriptional machinery. CBP can acetylate a broad range of these transcription factors, which, in most cases, potentiates transcription. Additionally, acetylation of histones by CBP facilitates gene transcription by providing an open chromatin structure (Blobel, 2000). Importantly, mice with CBP haploinsufficiency develop multilineage defects in hematopoietic differentiation and increased hematologic malignancies with age (Kung *et al.*, 2000), whereas conditional inactivation of CBP in murine T-cell precursors results in a high incidence of T-cell tumors (Kang-Decker *et al.*, 2004).

As ectopic expression of *HOXA* homeobox genes implicated in the pathogenesis of T-ALL also perturbs myeloerythroid differentiation in several model systems (Owens and Hawley, 2002), we reasoned that *TLX1* and *HOXA* oncogenes may act in part by targeting global regulatory circuits that impact cell proliferation and differentiation outcomes. In this regard, a number of oncogenic transcription factors have been observed to inhibit CBP activity in the context of cell differentiation arrest (Blobel, 2000). Among the best characterized examples are those that interfere with CBP-mediated acetylation of the transcription factor GATA-1, a key regulator of erythropoiesis (Blobel *et al.*, 1998; Hung *et al.*, 1999; Hong *et al.*, 2002). In the present work, we established a murine erythroid progenitor cell line, iEBHX1S-4, from ESC-derived embryoid bodies by conditional *TLX1* expression and we used this cell line to investigate the mechanism by which *TLX1* achieves differentiation arrest. The results suggest a mechanism by which sequestration of CBP by *TLX1* within particular subnuclear compartments might limit its access to critical acetylation substrates, such as GATA-1 in the case of erythroid differentiation.

Results

Upregulation of erythroid transcriptional networks in iEBHX1S-4 cells following release of the TLX1-mediated differentiation block

As described in the accompanying Supplementary Information, iEBHX1S-4 cells exhibit a proerythroblast-like

phenotype and require interleukin-3 plus stem cell factor for survival and proliferation (Supplementary Figure 1). Downregulation of *TLX1* expression releases the iEBHX1S-4 differentiation block, allowing erythropoietin-dependent acquisition of erythroid markers and hemoglobin synthesis (Supplementary Figure 2). Global gene expression profiles were determined by microarray profiling for iEBHX1S-4 cells at 0, 6, 12 and 24 h following doxycycline withdrawal. *TLX1* protein levels progressively decreased with a half-life of ~6 h, approaching basal levels that were below detection by Western blot analysis by 24 h (Supplementary Figure 2g; see Figure 2a). Unsupervised hierarchical clustering of the expression data created a condition tree that showed corresponding progressive changes in the iEBHX1S-4 transcriptome during the 24 h time course experiment (Figure 1a and b). Gene tree clustering revealed two major subtrees of genes whose transcript levels increased from 0 to 24 h (Figure 1c), which displayed nonrandom overlap ($P=0.011$) with a subset of genes induced upon restoration of GATA-1 activity during differentiation of ESC-derived GATA-1-null G1E erythroid cells (NCBI GEO Accession Number GDS568). We employed quantitative reverse transcription–polymerase chain reaction (qRT–PCR) to validate the expression pattern of a representative example from this set, *Ccne1* (cyclin E1), and selected examples of other induced genes that demonstrated different kinetics of upregulation, that is *Hba-x* (ζ -globin), *Hemgn* (hemogen) (Yang *et al.*, 2001) and *Apobec2* (Kostic and Shaw, 2000). The corresponding expression profiles for these genes are illustrated in Figure 1c.

To identify putative regulatory hierarchies downstream of *TLX1*, sets of conditionally regulated genes classified according to the Gene Ontology (GO) term ‘Transcription’ were subjected to bioinformatic promoter analysis. These included 85 gradually induced genes from the combined subtrees shown in Figure 1c (Supplementary Table 1), 46 representatives of the *Ccne1*-like profile ($\kappa>0.995$) (Supplementary Table 2), 42 representatives of the *Hba-x*-like profile ($\kappa>0.985$) (Supplementary Table 3), 31 representatives of the *Hemgn*-like profile ($\kappa>0.985$) (Supplementary Table 4) and 27 representatives of the *Apobec2*-like profile ($\kappa>0.975$) (Supplementary Table 5). For comparison, we included representatives from two subclusters of genes obtained by K-means clustering of the entire data set, whose transcripts were downregulated by 6 (44 representatives; Supplementary Table 6) or 12 h (45 representatives; Supplementary Table 7) following doxycycline withdrawal. A common feature of all of the transcription factors implicated through this analysis – GATA-1, KLF1, NF-Y, C/EBP and SCL – is that their transcriptional activity is regulated by CBP (see Supplementary Information). Given these observations, we hypothesized that *TLX1* might impede iEBHX1S-4 differentiation by interfering with CBP.

Impaired acetylation of GATA-1 in TLX1-expressing iEBHX1S-4 erythroid cells

We first determined whether the acetylation levels of GATA-1, an essential target for CBP-facilitated

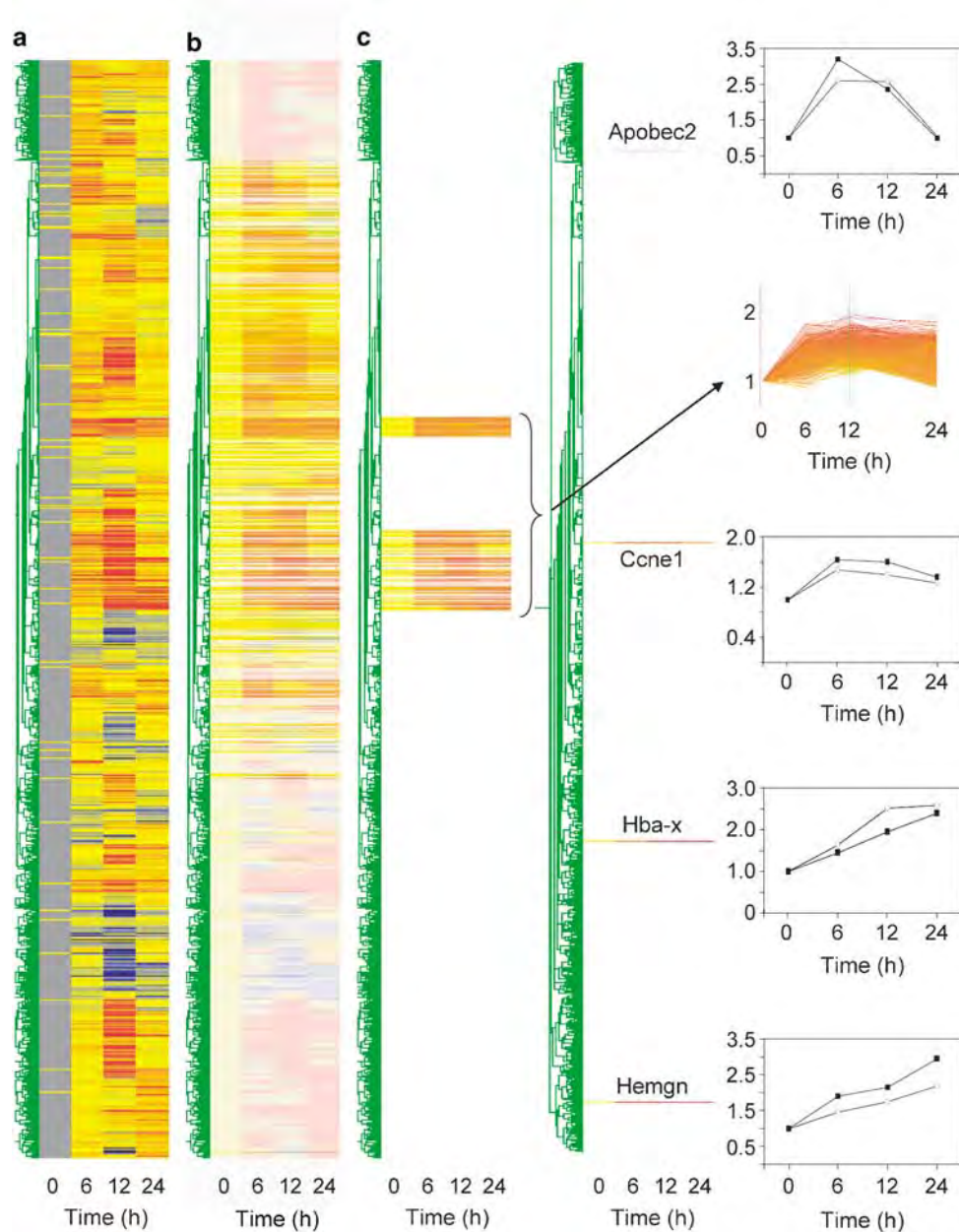


Figure 1 Overall analysis of the entire set of data showing expression changes in iEBHX1S-4 cells upon *TLX1* downregulation. Each microarray data set was normalized to the 50th percentile and then relative to corresponding signal intensities obtained for $t = 0$ h. (a) Condition and gene trees colored for significance. Blue corresponds to -3σ and red to 3σ . (b) Condition and gene trees colored for trust and expression levels. Blue corresponds to 0 and red to 2. Levels of trust increase with brightness. Graphs were generated using GeneSpring. (c) Selected subtrees for genes showing gradual increase during the observation period. Arrow indicates the corresponding expression profiles. Comparison of qRT-PCR (○) and microarray data (■) for selected induced transcripts (Ccne1/cyclin E1, Hba-x/ ζ -globin, Hemgn and Apobec2).

erythroid differentiation (Blobel *et al.*, 1998; Hong *et al.*, 2002), changed in iEBHX1S-4 cells upon release of the TLX1-mediated differentiation block. Indeed, following doxycycline withdrawal, the levels of acetylated GATA-1 increased (~ 2.5 -fold), whereas the levels of CBP-associated GATA-1 increased (~ 1.6 -fold) inversely proportional to the decreasing TLX1 protein levels during the 24 h time course experiment ($r = -0.90$ and -0.93 , respectively) (Figure 2a). Transcription

factor acetylation levels are the result of a dynamic equilibrium between acetyltransferases and deacetylases (Yang, 2004). In particular, GATA-1 has been demonstrated to associate with class I and class II histone deacetylase (HDAC) enzymes (Watanabe *et al.*, 2003; Rodriguez *et al.*, 2005). Because it was shown that treatment with the class I/II HDAC inhibitor, trichostatin A, markedly augmented acetylation of GATA-1 in transfected Cos 7 cells (Hernandez-Hernandez *et al.*,

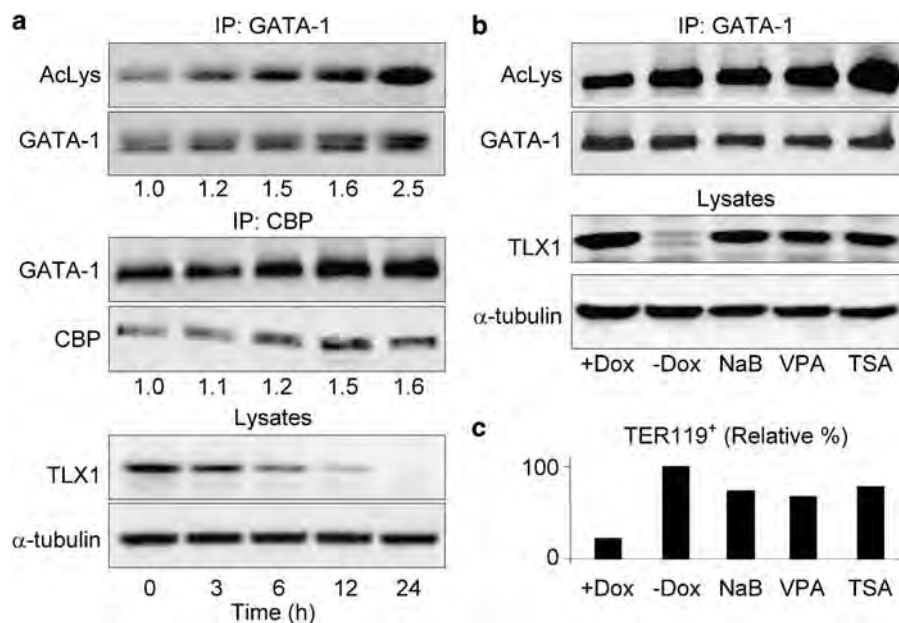


Figure 2 CBP interaction with GATA-1 in differentiating iEBHX1S-4 cells. (a) Change in GATA-1 acetylation and protein levels upon *TLX1* downregulation. The two top panels show Western blotting of anti-GATA-1 immunoprecipitates with anti-acetylated lysine or anti-GATA-1 antibodies. The ratios of acetylated GATA-1 to total GATA-1 are indicated. The two middle panels show Western blotting of anti-CBP immunoprecipitates with anti-GATA-1 or anti-CBP antibodies. The relative increase in CBP-associated GATA-1 levels is indicated. The two bottom panels show the corresponding decrease in total *TLX1* protein levels and an α -tubulin loading control. (b) Changes in GATA-1 acetylation following Dox withdrawal or treatment with HDAC inhibitors. The two top panels show Western blotting of anti-GATA-1 immunoprecipitates with anti-acetylated lysine or anti-GATA-1 antibodies. The two bottom panels show the corresponding *TLX1* protein levels and an α -tubulin loading control. + Dox indicates untreated iEBHX1S-4 cells cultured in the presence of 1 μ g/ml doxycycline; -Dox indicates cells grown without doxycycline for 24 h. Cells were treated with the indicated HDAC inhibitors for 24 h. Abbreviations: NaB, 1 mM sodium butyrate; VPA, 0.5 mM valproic acid; TSA, 50 nM trichostatin A. (c) The graph depicts levels of glycophorin A/TER119 surface antigen expression following doxycycline withdrawal or treatment with HDAC inhibitors. + Dox indicates untreated iEBHX1S-4 cells cultured in the presence of 1 μ g/ml doxycycline; -Dox indicates cells grown without doxycycline for 3 days. Cells were treated with the indicated HDAC inhibitors for 3 days. HDAC inhibitor abbreviations and concentrations as above. Glycophorin A/TER119 levels 3 days after doxycycline withdrawal were denoted as 100%.

2006), we next investigated whether class I/II HDAC inhibitor treatment would result in increased levels of acetylated GATA-1 in iEBHX1S-4 cells. We found that 24 h treatment of doxycycline-supplemented iEBHX1S-4 cell cultures with three specific class I/II HDAC inhibitors, sodium butyrate, valproic acid and trichostatin A, induced acetylation of GATA-1 comparable to the levels observed upon *TLX1* downregulation (Figure 2b). Based on these observations, we were interested in determining whether HDAC inhibitor treatment was sufficient to bypass the *TLX1*-mediated iEBHX1S-4 differentiation block (Yoshida *et al.*, 1987). Indeed, treatment of iEBHX1S-4 cells cultured in doxycycline-supplemented medium for 3 days with HDAC inhibitors resulted in considerable differentiation as reflected by upregulation of glycophorin A/TER119 expression, a target gene of the SCL-LMO2-GATA-1 complex (Lahlil *et al.*, 2004), with levels approaching that observed during the same period following doxycycline withdrawal (Figure 2c). These findings are consistent with the notion that insufficient GATA-1 acetylation levels are an important aspect of the *TLX1*-mediated differentiation arrest in iEBHX1S-4 cells.

TLX1 interaction with CBP in iEBHX1S-4 and 293T cells and targeting to heterochromatin

To obtain evidence in support of the possibility that *TLX1* might interfere with CBP function in iEBHX1S-4 cells, we performed coimmunoprecipitation experiments to determine whether *TLX1* was capable of physically associating with CBP *in vivo*. As shown in Figure 3a, *TLX1* coimmunoprecipitated with endogenous CBP from iEBHX1S-4 lysates. Coimmunoprecipitation of exogenous *TLX1* with exogenous mouse CBP from lysates of human 293T embryonic kidney cells cotransfected with expression vectors encoding *TLX1* (Owens *et al.*, 2003; Riz and Hawley, 2005) and mouse CBP (Chrivia *et al.*, 1993; Kwok *et al.*, 1994) was also demonstrated (Figure 3b, left panels). In addition, a monoclonal antibody directed against ectopically expressed FLAG epitope-tagged *TLX1* (Owens *et al.*, 2003) was shown to coimmunoprecipitate exogenous human CBP from lysates of 293T cells cotransfected with corresponding expression vectors in separate experiments (Figure 3b, right panels). We extended these studies by performing *in vitro* pull-down experiments with glutathione-S-transferase (GST)-*TLX1* fusion proteins. Both endogenous CBP from iEBHX1S-4

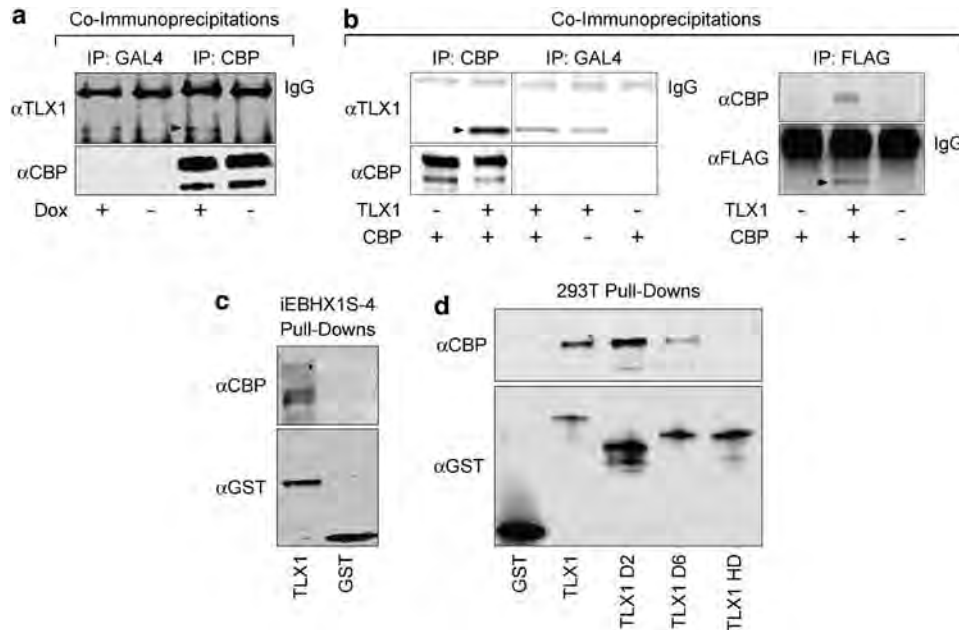


Figure 3 TLX1 interacts with CBP *in vivo* and *in vitro*. (a) Nuclear lysates of iEBHX1S-4 cells cultured in the presence of 1 μ g/ml doxycycline (+Dox) or grown without doxycycline for 3 days (–Dox) were immunoprecipitated with anti-CBP or anti-GAL4 (irrelevant control) antibodies followed by Western blot analysis with anti-TLX1 or anti-CBP antibodies. The TLX1 band is indicated by the arrowhead. (b) *Left* Whole-cell lysates of 293T cells transiently transfected with TLX1 or CBP expression vectors were immunoprecipitated with anti-CBP or anti-GAL4 (irrelevant control) antibodies followed by Western blot analysis with anti-TLX1 or anti-CBP antibodies. Under the conditions used, some nonspecific (background) immunoprecipitation of TLX1 was observed with the anti-GAL4 antibody. *Right* Whole-cell lysates of 293T cells transiently transfected with TLX1 (FLAG-tagged) or CBP expression vectors were immunoprecipitated with an anti-FLAG antibody followed by Western blot analysis with anti-CBP or anti-FLAG antibodies. TLX1 bands are indicated by the arrowheads. (c) iEBHX1S-4 nuclear lysates were incubated with immobilized GST–TLX1 fusion protein or with control GST beads and the bound proteins eluted and subjected to Western blot analysis with anti-CBP and anti-GST antibodies. The amount of eluate loaded to detect the GST–TLX1 fusion protein represents 0.5% of the amount loaded to detect CBP. (d) Nuclear lysates of 293T cells transiently transfected with a CBP expression vector were incubated with immobilized GST–TLX1 fusion proteins or with control GST beads and the bound proteins eluted and subjected to Western blot analysis with anti-CBP and anti-GST antibodies. The amount of each eluate loaded to detect GST–TLX1 fusion proteins represents 5% of the amount loaded to detect CBP. Abbreviations: TLX1, GST-FLAG-TLX1; TLX1 D2, GST-FLAG-TLX1 D2 (consisting of amino acids 98–330), GST-FLAG-TLX1 D6 (consisting of amino acids 2–260) and GST-FLAG-TLX1 HD (containing an internal deletion from amino acid 201 to amino acid 260).

lysates (Figure 3c) as well as ectopically expressed mouse CBP from 293T lysates (Figure 3d) bound to immobilized full-length GST–TLX1 fusion protein but not to control GST beads. Moreover, a GST–TLX1 fusion protein missing the homeodomain (TLX1 HD mutant) was incapable of coprecipitating exogenous mouse CBP from 293T lysates, whereas reduced binding was observed with a GST–TLX1 fusion protein containing a 70-amino-acid carboxy-terminal deletion (TLX1 D6 mutant), which truncated the TLX1 protein immediately after the homeodomain (Figure 3d). By comparison, coprecipitation of exogenous mouse CBP from 293T lysates with a GST–TLX1 fusion protein containing a 97-amino-acid amino-terminal deletion (TLX1 D2 mutant) was comparable to that achieved with the full-length GST–TLX1 fusion protein (Figure 3d). The combined results indicated that: (1) TLX1 is capable of interacting with endogenous and exogenous CBP under *in vivo* conditions; (2) *in vivo* TLX1–CBP complex formation did not depend on an erythroid lineage- or stage-specific nuclear structure or on erythroid-specific cofactors; (3) TLX1 directly interacts with CBP *in vitro* and (4) the homeodomain of TLX1 was required for

in vitro interaction with CBP, as was previously demonstrated for a number of clustered HOX proteins (Shen *et al.*, 2001).

We next investigated the intracellular distribution of TLX1 and CBP. iEBHX1S-4 cells grown in the presence or absence of doxycycline were fixed, immunolabeled with anti-TLX1 and/or anti-CBP antibodies, and examined by immunofluorescence staining and confocal laser scanning microscopy. As expected from previous findings (Chrivia *et al.*, 1993; Dear *et al.*, 1993; Owens *et al.*, 2003), TLX1 and CBP localized selectively within the nucleus. In the presence of doxycycline, significant colocalization of the two proteins was observed (Figure 4a, +Dox Merge). Because a recent publication reported that a proportion of TLX1 in human T-ALL cells unexpectedly localizes to heterochromatin domains (Heidari *et al.*, 2006), we were interested in examining whether TLX1 inhibition of CBP might result from the ‘intranuclear marshaling’ of CBP to heterochromatic regions (Schaufele *et al.*, 2001). Therefore, we next determined the intranuclear distribution of CBP in conditionally arrested iEBHX1S-4 cells with respect to heterochromatin markers. Lysine 9 methylation of

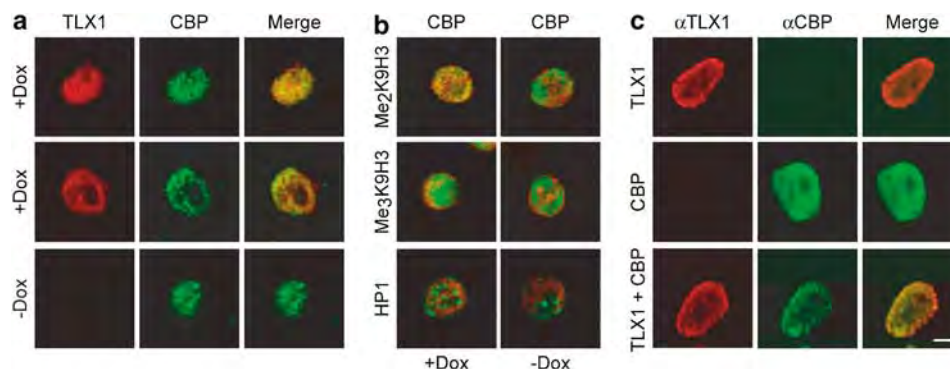


Figure 4 Partial colocalization of TLX1 and CBP in iEBHX1S-4 and 293T cells. **(a)** iEBHX1S-4 cells cultured in the presence of 1 μ g/ml doxycycline (+Dox) or grown without doxycycline for 18 h (–Dox) were labeled with anti-TLX1 (TLX1; Alexa Fluor 568, red) and anti-CBP (CBP; Alexa Fluor 488, green) antibodies and immunofluorescence staining was analysed by confocal laser scanning microscopy. The right panels show the merged green and red images at the same focal plane with overlapping regions of protein distribution appearing yellow. **(b)** iEBHX1S-4 cells cultured in the presence of 1 μ g/ml doxycycline (+Dox) or grown without doxycycline for 18 h (–Dox) were labeled with anti-CBP (CBP; Alexa Fluor 488, green) and either anti-dimethyl-histone H3 (Lys9) (Me₂K9H3; Alexa Fluor 568, red) or anti-trimethyl-histone H3 (Lys9) (Me₃K9H3; Alexa Fluor 568, red) antibodies, or with anti-CBP (CBP; Alexa Fluor 568, red) and anti-HP1 α (HP1; Alexa Fluor 488, green) antibodies and immunofluorescence staining was analysed by confocal laser scanning microscopy. The panels shown are the merged green and red images at the same focal plane. Overlapping distributions of CBP and dimethyl-histone H3 (Lys9) staining in +Dox cultures of iEBHX1S-4 cells appear yellow. **(c)** 293T cells transiently transfected with TLX1 and/or CBP expression vectors (indicated to the left of the panels) were labeled with anti-TLX1 (α TLX1; Alexa Fluor 568, red) and anti-CBP (α CBP; Alexa Fluor 488, green) antibodies, and immunofluorescence staining was analysed by confocal laser scanning microscopy. The right panels show the merged green and red images at the same focal plane with overlapping regions of protein distribution appearing yellow. Note that coexpression of TLX1 caused redistribution of a substantial fraction of CBP to the nuclear periphery (see Supplementary Figure 3 for details). Size bar, 10 μ m.

histone H3 (K9H3) is an epigenetic modification that has been correlated with both local and global repression of transcription, and a number of studies have suggested that the di- (Me₂K9H3) and trimethylation (Me₃K9H3) states of K9H3 largely reside in separate subnuclear compartments, possibly distinguishing facultative and constitutive heterochromatin, respectively (Guenatri *et al.*, 2004; Wu *et al.*, 2005). In addition, the α isoform of the non-histone adapter heterochromatin protein 1 (HP1 α) is frequently concentrated at Me₃K9H3-enriched heterochromatin (Guenatri *et al.*, 2004). In this regard, it was notable that costaining of iEBHX1S-4 cells grown in the presence of doxycycline with anti-CBP and anti-Me₂K9H3 antibodies revealed partially overlapping regions of fluorescence (Figure 4b, +Dox). In contrast, no overlap of CBP and Me₂K9H3 fluorescence was observed 18 h following doxycycline withdrawal (Figure 4b, –Dox), indicating exit of CBP from this subnuclear compartment concomitant with TLX1 downregulation. By comparison, no overlap of the CBP distribution pattern with Me₃K9H3 or with HP1 α was revealed by immunofluorescence confocal microscopy of iEBHX1S-4 cells similarly cultured in the presence or absence of doxycycline (Figure 4b). These results suggested that TLX1 might inhibit CBP function in iEBHX1S-4 cells by sequestering a sub-population of the protein in particular subnuclear compartments, including those associated with heterochromatin domains enriched in Me₂K9H3.

In light of these observations, we were interested in directly studying the effect of TLX1 expression on the intranuclear distribution of CBP. Therefore, we transiently transfected 293T cells with the FLAG-tagged TLX1 and/or mouse CBP expression vectors and

examined their intranuclear locations by immunofluorescence staining and confocal laser scanning microscopy (Figure 4c; Supplementary Figure 3). Under these experimental conditions, TLX1 was preferentially located at the nuclear periphery, whereas in the absence of TLX1, CBP was distributed throughout the nucleus. Quantitative image analysis (Supplementary Figure 3) revealed that there was a statistically significant difference between the distribution of TLX1 in the peripheral versus the central region of the nucleus ($P=0.024$) but not in the case of CBP ($P=0.328$, peripheral versus central localization). However, when TLX1 was coexpressed with CBP, a substantial fraction of CBP exhibited a striking redistribution to the nuclear periphery ($P=0.001$, peripheral versus central localization), colocalizing with TLX1 (Pearson correlation coefficient, $r=0.672$). These results provided direct evidence for the recruitment of CBP to subnuclear compartments occupied by coexpressed TLX1.

Discussion

We inferred from previous work in various murine model systems that *TLX1* functions in human leukemia etiology at least in part by disrupting hematopoietic differentiation programs (Hawley *et al.*, 1994, 1997; Keller *et al.*, 1998; Owens *et al.*, 2003, 2006). The collective observations thus raised the possibility that TLX1 might interfere with hematopoietic differentiation pathways by interacting with shared signaling components or transcriptional coregulators. To gain a better understanding of the underlying mechanism of the TLX1-mediated differentiation block, we generated the

factor-dependent iEBHX1S-4 progenitor cell line by conditional immortalization with doxycycline-inducible *TLX1* expression. We then performed genome-wide expression profiling of iEBHX1S-4 cells released from the differentiation block at early time points following doxycycline withdrawal. A key feature of our experimental design was the bioinformatic analysis of functionally related sets of genes exhibiting similar expression profiles following TLX1 extinction. This analysis revealed coordinated upregulation of erythroid transcriptional networks integrated by the acetyltransferase co-activator CBP. Among erythroid-lineage transcription factor targets of CBP, previous work had highlighted CBP acetylation of GATA-1 as being essential for erythroid differentiation (Blobel *et al.*, 1998; Hung *et al.*, 1999; Hong *et al.*, 2002). Accordingly, we found immediate increases in the levels of CBP-associated GATA-1 as well as the acetylated form of GATA-1 upon TLX1 downregulation, whereas class I/II HDAC inhibitor treatment of conditionally arrested iEBHX1S-4 cells stimulated GATA-1 acetylation and differentiation (Yoshida *et al.*, 1987; Watamoto *et al.*, 2003). We subsequently demonstrated by coimmunoprecipitation experiments and GST pull-down assays that TLX1 binds to CBP *in vivo* and *in vitro*, and we provided evidence that the homeodomain of TLX1 is required for its direct interaction with CBP *in vitro*. We also showed by confocal laser microscopy that CBP partially colocalizes with TLX1 and Me₂K9H3-marked heterochromatin in iEBHX1S-4 cells, relocating from these heterochromatic regions concomitant with TLX1 downregulation. Further, we documented that coexpression of TLX1 with CBP in a heterologous cell line (293T cells) resulted in the redistribution of its intranuclear location. The combined results presented here can therefore be interpreted to suggest a mechanism by which TLX1 modulates CBP function by binding and recruiting it to particular subnuclear compartments, including those organized into repressive chromatin domains (Schaufele *et al.*, 2001; Heidari *et al.*, 2006).

Transforming viral proteins such as adenovirus E1A, which force cells into Sphase, target CBP as well as the retinoblastoma (Rb) protein (Blobel, 2000; Helt and Galloway, 2003). We previously showed that TLX1 regulated multiple G₁/S transcriptional networks in *TLX1*⁺ human T-ALL cell lines by inhibiting Rb function (Riz and Hawley, 2005). Whereas it is clear that the adenovirus E1A oncoprotein represses Rb activity, opposing effects of E1A on CBP activity have been reported (Ait-Si-Ali *et al.*, 2000). Notably, although E1A interferes with CBP-mediated acetylation of GATA-1 (Blobel *et al.*, 1998; Hung *et al.*, 1999), E1A modulates expression of certain cell cycle-related genes such as the proliferating cell nuclear antigen in part by disrupting CBP interaction with other transcriptional regulators (Karuppayil *et al.*, 1998). Thus, the current findings leave open the possibility that TLX1 may also redirect as well as inhibit CBP-facilitated differentiation signals, converting them into proliferative responses.

CBP and the closely related p300 protein function as global coregulators of transcription, purportedly

interacting physically or functionally with over 300 proteins (Kasper *et al.*, 2006). It is not surprising therefore that many developmental pathways culminate in interactions that involve CBP. In particular, a full complement of CBP is required for normal differentiation along multiple hematopoietic lineages (Kung *et al.*, 2000; Kasper *et al.*, 2006). The current studies using the novel iEBHX1S-4 erythroid progenitor cell model suggest that the mechanism by which TLX1 contributes to erythroid differentiation arrest occurs in a manner analogous to that for several other oncoproteins (Blobel *et al.*, 1998; Hung *et al.*, 1999; Hong *et al.*, 2002). In this regard, it is worth noting that subversion of erythroid transcriptional networks is observed in human T-ALL cases in connection with the SCL and LMO2 transcription factors, which normally form a DNA-binding complex containing GATA-1 in erythroid cells (Wadman *et al.*, 1997). Similar to *TLX1* (Owens *et al.*, 2006), enforced expression of *SCL* or *LMO2* in thymocyte precursors causes deregulation of the transition checkpoint from the CD4[−] CD8[−] double-negative to CD4⁺ CD8⁺ double-positive stages of T-cell development (Larson *et al.*, 1995; Herblot *et al.*, 2000), a consequence mimicked by attenuating CBP activity during thymocyte development (Kasper *et al.*, 2006). In the case of SCL, both activator and repressor functions have been ascribed to multiprotein complexes, exerted through SCL association with CBP and other protein partners (Huang *et al.*, 1999; Schuh *et al.*, 2005). Of further interest is the recent observation that SCL associates with heterochromatin domains and mediates regional transcriptional repression by a chromatin remodeling mechanism that is sensitive to the class I/II HDAC inhibitor trichostatin A (Wen *et al.*, 2005).

Modulation of CBP function in the context of differentiation arrest is also a recurring theme in human acute myeloid leukemia, with chromosomal translocations frequently targeting CBP directly or the resulting fusion proteins – for example, MOZ–TIF2, AML1–ETO – shown to interact with CBP (Deguchi *et al.*, 2003; Iyer *et al.*, 2004; Choi *et al.*, 2006). It is noteworthy, for example, that interaction with CBP is necessary for immortalization of murine myeloid progenitors by the MOZ–TIF2 oncoprotein (Deguchi *et al.*, 2003). Interaction with CBP has also been proposed to play a role in the immortalization of murine myeloid progenitors by the E2A–PBX1 fusion oncoprotein of human pre-B-cell ALL (Kamps and Wright, 1994; Bayly *et al.*, 2004). Of particular relevance to the current study is the demonstration that the differentiation of certain murine myeloid progenitor cell lines conditionally immortalized by *E2A-PBX1* could be arrested by ectopic expression of a variety of oncogenes, including *AML1-ETO*, *HOXA7* and *HOXA9* as well as other *HOX* genes (Sykes and Kamps, 2001). The accumulated data, considered together with previous observations that many HOX proteins were found to interact with CBP, commonly via the homeodomain (Shen *et al.*, 2001), suggest a shared indirect mechanism of hematopoietic cell differentiation arrest mediated by these homeodomain-containing transcription factors. In view of the recent appreciation

of deregulated *HOXA* homeobox gene expression in human T-ALL and the finding that *HOXA*-translocated samples could be grouped together with *TLX1*⁺ cases based on genome-wide expression analysis (Soulier *et al.*, 2005), it is tempting to speculate that modulation of CBP function may contribute to T-ALL evoked by the TLX1 and HOXA homeodomain proteins (Kang-Decker *et al.*, 2004).

Materials and methods

iEBHX1S-4 erythroid progenitor cell line derivation

The ploxTLX1 targeting plasmid was electroporated into the doxycycline-inducible ESC line Ainv15 and selected for G418 resistance as described (Kyba *et al.*, 2002). Embryoid body formation, and iEBHX1S-4 progenitor cell line derivation and characterization were essentially as described (Keller *et al.*, 1998; Kyba *et al.*, 2002). See Supplementary Information for details.

Microarray profiling

Microarray profiling was performed in The George Washington University Medical Center Genomics Core Facility

References

- Ait-Si-Ali S, Polesskaya A, Filleur S, Ferreira R, Duquet A, Robin P *et al.* (2000). CBP/p300 histone acetyl-transferase activity is important for the G1/S transition. *Oncogene* **19**: 2430–2437.
- Akimov SS, Ramezani A, Hawley TS, Hawley RG. (2005). Bypass of senescence, immortalization, and transformation of human hematopoietic progenitor cells. *Stem Cells* **23**: 1423–1433.
- Allen TD, Zhu Y-X, Hawley TS, Hawley RG. (2000). TALE homeoproteins as HOX11-interacting partners in T-cell leukemia. *Leuk Lymphoma* **39**: 241–256.
- Bayly R, Chuen L, Currie RA, Hyndman BD, Casselman R, Blobel GA *et al.* (2004). E2A-PBX1 interacts directly with the KIX domain of CBP/p300 in the induction of proliferation in primary hematopoietic cells. *J Biol Chem* **279**: 55362–55371.
- Berger LC, Hawley RG. (1997). Interferon- β interrupts interleukin-6-dependent signaling events in myeloma cells. *Blood* **89**: 261–271.
- Blobel GA. (2000). CREB-binding protein and p300: molecular integrators of hematopoietic transcription. *Blood* **95**: 745–755.
- Blobel GA, Nakajima T, Eckner R, Montminy M, Orkin SH. (1998). CREB-binding protein cooperates with transcription factor GATA-1 and is required for erythroid differentiation. *Proc Natl Acad Sci USA* **95**: 2061–2066.
- Choi Y, Elagib KE, Delehanty LL, Goldfarb AN. (2006). Erythroid inhibition by the leukemic fusion AML1-ETO is associated with impaired acetylation of the major erythroid transcription factor GATA-1. *Cancer Res* **66**: 2990–2996.
- Chrivia JC, Kwok RP, Lamb N, Hagiwara M, Montminy MR, Goodman RH. (1993). Phosphorylated CREB binds specifically to the nuclear protein CBP. *Nature* **365**: 855–859.
- Dear TN, Sanchez-Garcia I, Rabbitts TH. (1993). The *HOX11* gene encodes a DNA-binding nuclear transcription factor belonging to a distinct family of homeobox genes. *Proc Natl Acad Sci USA* **90**: 4431–4435.
- Deguchi K, Ayton PM, Carapeti M, Kutok JL, Snyder CS, Williams IR *et al.* (2003). MOZ-TIF2-induced acute myeloid leukemia requires the MOZ nucleosome binding motif and TIF2-mediated recruitment of CBP. *Cancer Cell* **3**: 259–271.
- Greene WK, Bahn S, Masson N, Rabbitts TH. (1998). The T-cell oncogenic protein HOX11 activates *Aldh1* expression in NIH 3T3 cells but represses its expression in mouse spleen development. *Mol Cell Biol* **18**: 7030–7037.
- Guenatri M, Bailly D, Maison C, Almouzni G. (2004). Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *J Cell Biol* **166**: 493–505.
- Hawley RG, Fong AZC, Lu M, Hawley TS. (1994). The *HOX11* homeobox-containing gene of human leukemia immortalizes murine hematopoietic precursors. *Oncogene* **9**: 1–12.
- Hawley RG, Fong AZC, Reis MD, Zhang N, Lu M, Hawley TS. (1997). Transforming function of the *HOX11/TCL3* homeobox gene. *Cancer Res* **57**: 337–345.
- Heidari M, Rice KL, Phillips JK, Kees UR, Greene WK. (2006). The nuclear oncoprotein TLX1/HOX11 associates with pericentromeric satellite 2 DNA in leukemic T-cells. *Leukemia* **20**: 304–312.
- Helt AM, Galloway DA. (2003). Mechanisms by which DNA tumor virus oncoproteins target the Rb family of pocket proteins. *Carcinogenesis* **24**: 159–169.
- Herblot S, Steff AM, Hugo P, Aplan PD, Hoang T. (2000). SCL and LMO1 alter thymocyte differentiation: inhibition of E2A-HEB function and pre-T α chain expression. *Nat Immunol* **1**: 138–144.
- Hernandez-Hernandez A, Ray P, Litos G, Ciro M, Ottolenghi S, Beug H *et al.* (2006). Acetylation and MAPK phosphorylation cooperate to regulate the degradation of active GATA-1. *EMBO J* **25**: 3264–3274.
- Hong W, Kim AY, Ky S, Rakowski C, Seo SB, Chakravarti D *et al.* (2002). Inhibition of CBP-mediated protein acetylation by the Ets family oncoprotein PU 1. *Mol Cell Biol* **22**: 3729–3743.
- Huang S, Qiu Y, Stein RW, Brandt SJ. (1999). p300 functions as a transcriptional coactivator for the TAL1/SCL oncoprotein. *Oncogene* **18**: 4958–4967.

Immunoprecipitations, GST pull-downs and Western blotting

Immunoprecipitations, GST pull-downs and Western blotting were performed essentially as described previously (Berger and Hawley, 1997; Owens *et al.*, 2003; Akimov *et al.*, 2005; Riz and Hawley, 2005).

Confocal laser scanning microscopy and image analysis

Confocal images were acquired using the $\times 60$ oil immersion objective of a Bio-Rad MRC-1024 confocal laser scanning microscope equipped with an argon-krypton ion laser and LaserSharp 2000 software (Carl Zeiss MicroImaging Inc., Thornwood, NY, USA) and were analysed using Image-Pro Plus (Media Cybernetics, Silver Spring, MD, USA) as described previously (Popratiloff *et al.*, 2003) as detailed in Supplementary Information.

- Hung HL, Lau J, Kim AY, Weiss MJ, Blobel GA. (1999). CREB-binding protein acetylates hematopoietic transcription factor GATA-1 at functionally important sites. *Mol Cell Biol* **19**: 3496–3505.
- Iyer NG, Ozdag H, Caldas C. (2004). p300/CBP and cancer. *Oncogene* **23**: 4225–4231.
- Kamps MP, Wright DD. (1994). Oncoprotein E2A-Pbx1 immortalizes a myeloid progenitor in primary marrow cultures without abrogating its factor-dependence. *Oncogene* **9**: 3159–3166.
- Kang-Decker N, Tong C, Boussouar F, Baker DJ, Xu W, Leontovich AA *et al.* (2004). Loss of CBP causes T cell lymphomagenesis in synergy with p27Kip1 insufficiency. *Cancer Cell* **5**: 177–189.
- Karuppayil SM, Moran E, Das GM. (1998). Differential regulation of p53-dependent and -independent proliferating cell nuclear antigen gene transcription by 12 O⁶-E1A oncoprotein requires CBP. *J Biol Chem* **273**: 17303–17306.
- Kasper LH, Fukuyama T, Biesen MA, Boussouar F, Tong C, de Pauw A *et al.* (2006). Conditional knockout mice reveal distinct functions for the global transcriptional coactivators CBP and p300 in T-cell development. *Mol Cell Biol* **26**: 789–809.
- Kawabe T, Muslin AJ, Korsmeyer SJ. (1997). HOX11 interacts with protein phosphatases PP2A and PP1 and disrupts a G₂/M cell-cycle checkpoint. *Nature* **385**: 454–458.
- Keller G, Wall C, Fong AZC, Hawley TS, Hawley RG. (1998). Overexpression of HOX11 leads to the immortalization of embryonic precursors with both primitive and definitive hematopoietic potential. *Blood* **92**: 877–887.
- Kostic C, Shaw PH. (2000). Isolation and characterization of sixteen novel p53 response genes. *Oncogene* **19**: 3978–3987.
- Krasnoselskaya-Riz I, Spruill A, Chen YW, Schuster D, Teslovich T, Baker C *et al.* (2002). Nuclear factor 90 mediates activation of the cellular antiviral expression cascade. *AIDS Res Hum Retroviruses* **18**: 591–604.
- Kung AL, Rebel VI, Bronson RT, Ch'ng LE, Sieff CA, Livingston DM *et al.* (2000). Gene dose-dependent control of hematopoiesis and hematologic tumor suppression by CBP. *Genes Dev* **14**: 272–277.
- Kwok RP, Lundblad JR, Chrivia JC, Richards JP, Bachinger HP, Brennan RG *et al.* (1994). Nuclear protein CBP is a coactivator for the transcription factor CREB. *Nature* **370**: 223–226.
- Kyba M, Perlingeiro RCR, Daley GQ. (2002). HoxB4 confers definitive lymphoid-myeloid engraftment potential on embryonic stem cell and yolk sac hematopoietic precursors. *Cell* **109**: 29–37.
- Lahlil R, Lecuyer E, Herblot S, Hoang T. (2004). SCL assembles a multifactorial complex that determines glycoporphin A expression. *Mol Cell Biol* **24**: 1439–1452.
- Larson RC, Osada H, Larson TA, Lavenir I, Rabbitts TH. (1995). The oncogenic LIM protein Rbtl2 causes thymic developmental aberrations that precede malignancy in transgenic mice. *Oncogene* **11**: 853–862.
- Owens BM, Hawley RG. (2002). *HOX* and non-*HOX* homeobox genes in leukemic hematopoiesis. *Stem Cells* **20**: 364–379.
- Owens BM, Hawley TS, Spain LM, Kerkel KA, Hawley RG. (2006). *TLX1/HOX11*-mediated disruption of primary thymocyte differentiation prior to the CD4⁺CD8⁺ double-positive stage. *Br J Haematol* **132**: 216–229.
- Owens BM, Zhu YX, Suen TC, Wang PX, Greenblatt JF, Goss PE *et al.* (2003). Specific homeodomain-DNA interactions are required for HOX11-mediated transformation. *Blood* **101**: 4966–4974.
- Popratiloff A, Giaume C, Peusner KD. (2003). Developmental change in expression and subcellular localization of two shaker-related potassium channel proteins (Kv1.1 and Kv1.2) in the chick tangential vestibular nucleus. *J Comp Neurol* **461**: 466–482.
- Riz I, Hawley RG. (2005). G₁/S transcriptional networks modulated by the *HOX11/TLX1* oncogene of T-cell acute lymphoblastic leukemia. *Oncogene* **24**: 5561–5575.
- Rodriguez P, Bonte E, Krijgsvelde J, Kolodziej KE, Guyot B, Heck AJ *et al.* (2005). GATA-1 forms distinct activating and repressive complexes in erythroid cells. *EMBO J* **24**: 2354–2366.
- Schaufele F, Enwright III JF, Wang X, Teoh C, Srihari R, Erickson R *et al.* (2001). CCAAT/enhancer binding protein α assembles essential cooperating factors in common subnuclear domains. *Mol Endocrinol* **15**: 1665–1676.
- Schuh AH, Tipping AJ, Clark AJ, Hamlett I, Guyot B, Iborra FJ *et al.* (2005). ETO-2 associates with SCL in erythroid cells and megakaryocytes and provides repressor functions in erythropoiesis. *Mol Cell Biol* **25**: 10235–10250.
- Shen WF, Krishnan K, Lawrence HJ, Largman C. (2001). The HOX homeodomain proteins block CBP histone acetyltransferase activity. *Mol Cell Biol* **21**: 7509–7522.
- Soulier J, Clappier E, Cayuela JM, Regnault A, Garcia-Peydro M, Dombret H *et al.* (2005). *HOXA* genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood* **106**: 274–286.
- Sykes DB, Kamps MP. (2001). Estrogen-dependent E2a/Pbx1 myeloid cell lines exhibit conditional differentiation that can be arrested by other leukemic oncoproteins. *Blood* **98**: 2308–2318.
- Wadman IA, Osada H, Grutz GG, Agulnick AD, Westphal H, Forster A *et al.* (1997). The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J* **16**: 3145–3157.
- Watanoto K, Towatari M, Ozawa Y, Miyata Y, Okamoto M, Abe A *et al.* (2003). Altered interaction of HDAC5 with GATA-1 during MEL cell differentiation. *Oncogene* **22**: 9176–9184.
- Wen J, Huang S, Pack SD, Yu X, Brandt SJ, Noguchi CT. (2005). Tal1/SCL binding to pericentromeric DNA represses transcription. *J Biol Chem* **280**: 12956–12966.
- Wu R, Terry AV, Singh PB, Gilbert DM. (2005). Differential subnuclear localization and replication timing of histone H3 lysine 9 methylation states. *Mol Biol Cell* **16**: 2872–2881.
- Yang LV, Nicholson RH, Kaplan J, Galy A, Li L. (2001). Hemogen is a novel nuclear factor specifically expressed in mouse hematopoietic development and its human homologue EDAG maps to chromosome 9q22, a region containing breakpoints of hematological neoplasms. *Mech Dev* **104**: 105–111.
- Yang XJ. (2004). Lysine acetylation and the bromodomain: a new partnership for signaling. *BioEssays* **26**: 1076–1087.
- Yoshida M, Nomura S, Beppu T. (1987). Effects of trichostatin on differentiation of murine erythroleukemia cells. *Cancer Res* **47**: 3688–3691.
- Zhang N, Shen W, Hawley RG, Lu M. (1999). HOX11 interacts with CTF1 and mediates hematopoietic precursor cell immortalization. *Oncogene* **18**: 2273–2279.

Supplementary Information accompanies the paper on the Oncogene website (<http://www.nature.com/onc>).

Database Updates

The Histone Database: an integrated resource for histones and histone fold-containing proteins

Leonardo Mariño-Ramírez^{1,*}, Kevin M. Levine¹, Mario Morales², Suiyuan Zhang³, R. Travis Moreland³, Andreas D. Baxevanis³ and David Landsman¹

¹Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, MSC 6075, Bethesda, MD 20894-6075, USA, ²Polytechnic Institute of New York University, Six MetroTech Center, Brooklyn, NY 11201, USA and ³Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Building 50, Room 5222, Bethesda, MD 20892-8002

*Corresponding author: Tel: +1 301 402 3708; Fax: +1 301 480 2288; Email: marino@ncbi.nlm.nih.gov

Submitted 6 July 2011; Revised 30 September 2011; Accepted 3 October 2011

Eukaryotic chromatin is composed of DNA and protein components—core histones—that act to compactly pack the DNA into nucleosomes, the fundamental building blocks of chromatin. These nucleosomes are connected to adjacent nucleosomes by linker histones. Nucleosomes are highly dynamic and, through various core histone post-translational modifications and incorporation of diverse histone variants, can serve as epigenetic marks to control processes such as gene expression and recombination. The Histone Sequence Database is a curated collection of sequences and structures of histones and non-histone proteins containing histone folds, assembled from major public databases. Here, we report a substantial increase in the number of sequences and taxonomic coverage for histone and histone fold-containing proteins available in the database. Additionally, the database now contains an expanded dataset that includes archaeal histone sequences. The database also provides comprehensive multiple sequence alignments for each of the four core histones (H2A, H2B, H3 and H4), the linker histones (H1/H5) and the archaeal histones. The database also includes current information on solved histone fold-containing structures. The Histone Sequence Database is an inclusive resource for the analysis of chromatin structure and function focused on histones and histone fold-containing proteins.

Database URL: The Histone Sequence Database is freely available and can be accessed at <http://research.nhgri.nih.gov/histones/>.

Introduction

Histones play central roles in both chromatin organization and gene regulation, as they constitute the fundamental protein units of the nucleosome (1). The nucleosome consists of DNA wrapped around an octameric core histone complex, composed of a central H3–H4 tetramer and two adjacent H2A–H2B dimers; the nucleosome is commonly identified as the first order of compaction of eukaryotic chromatin (2). Core histone genes also display conserved expression patterns that show periodic expression across the eukaryotic cell cycle, with a pronounced peak during

S-phase (3). This allows for histone proteins to be produced at the same time DNA is being synthesized. Thus, the histone proteins can be readily assembled into nucleosomes and then compacted into chromatin.

Core histones are highly conserved across eukaryotes in terms of sequence and structure. Despite overall sequence conservation, extensive histone tail post-translational modifications, in addition to histone variants present during development, contribute to epigenetic mechanisms that signal transcriptional activation, repression and recombination events. Histone proteins and their variants have an

essential function in gene regulation (4–6). Nucleosomes are disassembled at transcriptionally active promoters via histone post-translational modifications (7), specific histone variants are known to mark active promoters and regulatory regions (8), and other variants are involved in the transition between transcriptionally active or silent chromatin (9). In recent years, much progress has been made toward genome-wide profiling of chromatin modifications (10), where histones play critical roles in defining the overall structure and function of chromatin and, by extension, in gene regulation.

The histone fold is a common structural motif shared by each of the four core histones, which mediates interactions between the individual core histones. The histone fold is structurally composed of three α -helices connected by two loops, and this overall architecture allows for heterodimeric interactions between core histones (11). Interestingly, even though each individual histone protein family is highly conserved, the histone fold is not conserved at the sequence level but, rather, at the structural level (12). Higher-resolution crystal structure of the nucleosome core particle has demonstrated detailed structures of the histone folds in each of the histones (13, 14). The DNA wrapped around each nucleosome is held in place by linker histones (called H1, or H5 in avian species). The linker histones, which do not contain the histone fold motif and have a different evolutionary origin from the core histones (15), are critical to chromatin higher-order compaction and facilitate internucleosomal interactions (16). In addition, H1 variants have been shown to be involved in the regulation of developmental genes (17). The overall structural state of chromatin controls DNA replication, recombination and gene expression, with histones playing critical roles during these processes (18).

Interestingly, despite the conservation of core histone gene expression patterns, the regulatory machinery that controls core histone gene expression has changed greatly among eukaryotic evolutionary lineages. Specifically, the identity of the core histone gene *cis*-regulatory sequence motifs and the protein factors that bind these motifs are distinct for the yeast *Saccharomyces cerevisiae*, as well as for other fungi, plants, insects and mammals (19). Therefore, different species have developed unique gene regulatory mechanisms for core histone genes that converge in the same gene expression phenotype, high expression levels specifically during S phase, concomitant with DNA replication.

Although the core histones are among the most slowly evolving eukaryotic proteins, members of the histone H2A and H3 families have diversified extensively, assuming specialized roles in DNA repair, gene silencing, gene expression and centromere function (5, 6). Interestingly, the centromere H3 variant appears to form tetrameric nucleosomes that induce positive supercoils, and these specialized

‘centromeric nucleosomes’ have been proposed as the epigenetic inheritance mechanism for centromeres (20).

The histone fold motif—common to all core histones—has also been found in a variety of non-histone proteins. The large majority of these non-histone proteins are localized in the nucleus and their functions are related to DNA metabolism; they include nuclear factor Y (NF-Y) and the TFIIB transcription factors (12). A few histone fold-containing proteins localized in the cytoplasm include the Ras activator Son of Sevenless (SOS) (21): SOS1 is localized primarily in the nucleus and SOS2 localized in the cytoplasm (22). Huntingtin interacting protein M (CXorf27) also contains a histone fold and is localized in the cytoplasm. We hypothesize that histone folds in cytoplasm-localized proteins are used to mediate protein–protein interactions.

Given the central role of histones and related proteins in a wide variety of critical cellular functions, we feel the need to continue to provide a centralized, curated source of important information on these proteins to the biomedical community. To this end, the Histone Sequence Database represents an organized collection of all histones and histone fold-containing proteins (23). The information presented in this Database includes a list of published three-dimensional structures for histones and histone fold-containing proteins, as well as manually curated multiple sequence alignments for each histone family.

Database and software

Data tables

The Histone Sequence Database, which has been developed and expanded significantly since its last release (23), has three tables stored in a relational database schema using Oracle 10g (Figure 1). The HISTONES table stores information about the histone category, its accession, the sequence string, the submitting database, as well as NCBI’s taxonomic information on the sequence. The ORGANISM table contains detailed taxonomic information for the sequences contained in the Histone Sequence Database. The STRUCTURES table stores information on the experimentally determined structures of proteins contained in the database, including the method of determination (i.e. X-ray crystallography or NMR spectroscopy).

Software

The Histone Sequence Database uses Common Gateway Interfaces (CGIs) written in the Perl programming language that communicate with the relational database software. The connectivity between the CGIs and the Oracle 10g Relational Database Management Software was implemented using Perl’s Database Interface (DBI) and the Oracle database driver for the DBI module (DBD::Oracle), available through the Comprehensive Perl Archive Network

(CPAN; <http://search.cpan.org/>). The use of object-oriented design methodologies and Perl modules that are both open source and developed in-house allows for flexibility and scalability. The Web pages displaying data, such as the summary of contents, non-redundant sets, and search pages are dynamically generated using CGI. Comments concerning the Web front-end are welcomed and encouraged.

Data sources and histone protein identification

The protein databases searched for the update and curation of the Histone Sequence Database were the NCBI non-redundant (nr) database (18 November 2010); nr includes sequences of all non-redundant GenBank CDS translations (24), as well as the sequences of RefSeq proteins, sequences of structures represented in the Protein Data Bank (PDB) (25), and sequences from UniProtKB/Swiss-Prot (26), the Protein Information Resource (PIR) (27), and the Protein Research Foundation (PRF) (<http://www.prf.or.jp/index-e.html>). The collection of histones was extended and revised, using the HMMER3 software package (28). We constructed hidden Markov models (HMMs) for each of the four core histones and the linker histone H1 from the alignments generated in the last release of the Histone Database. Additional HMMs were generated for archaeal histones (29) and bacterial proteins that contain a

histone-likefold (30); only the protein entries that have a complete domain hit with an $E < 0.01$ are collected for further analysis. For each histone family, multiple sequence alignments were generated using MUSCLE (31). The alignments that are manually curated to include proteins with complete folds are also available in PDF format and are color-coded to allow easy identification of amino acid variants. The Histone Database uses a color scheme designed to highlight the specific amino acid differences that a particular group of sequences may have inside the core or linker histone alignments by coloring amino acids with similar physicochemical properties differently. A summary table of the number of sequences found grouped by family and species represented in the database is provided (Table 1).

Identification of histone fold-containing proteins

Histone fold-containing proteins were identified using a different search strategy. We used the sequences from each of the four core histone MUSCLE alignments (H2A, H2B, H3 and H4) as seeds for PSI-BLAST (32) searches. The PSI-BLAST searches were run to convergence with an E -value inclusion threshold of 0.01; the core histone seeds were excluded from the final list of histone fold-containing proteins. Additionally, related structures were identified using NCBI's VAST-related structures searches (33, 34), in an effort to identify more distant histone fold-containing proteins that could not be identified through PSI-BLAST searches. Using this strategy, we were able to identify a total of 2180 histone fold-containing proteins.

Results and Discussion

The computational approach presented here has identified proteins throughout a wider evolutionary spread of genomes. Currently, the Histone Database contains entries that represent a total of 7356 unique NCBI taxonomic identifiers, which correspond to approximately the same number of organisms. The sequences of core histones, linker histones and archeal histones are available in FASTA format.

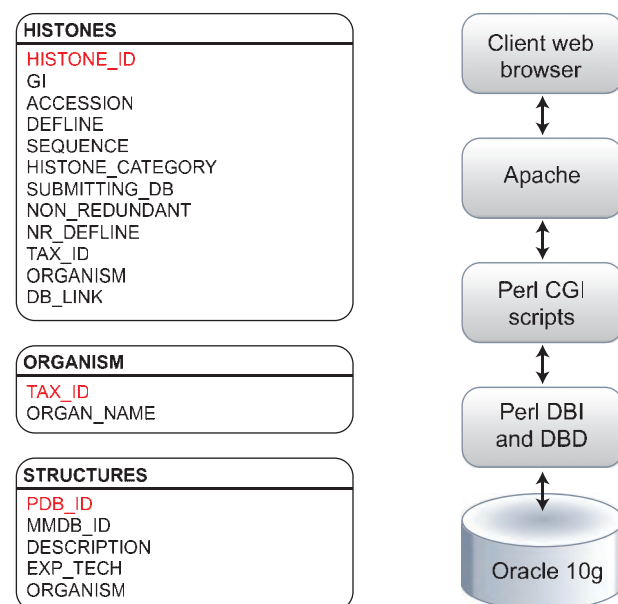


Figure 1. Histone Database data model. The Histone Database stores selected manually curated information from GenBank records. The information stored as part of each record includes the GenBank unique identifier (GI), accession number, definition line, sequence string, histone class, database source, NCBI taxonomic identifier and organism name. The database front-end is written in Perl, the data is stored in an Oracle 10g relational database, and data is retrieved using Perl DBI and DBD libraries.

Table 1. Histone Database content

Core histone profile	Number of unique sequences	Increase since last update (%) (23)	Number of unique taxonomic identifiers
H1/H5	591	138.3	156
H2A	1016	214.6	331
H2B	755	161.2	308
H3	2287	476.1	7096
H4	341	181.8	344
Archaeal	182	–	89

A

genome.gov
National Human Genome Research Institute
National Institutes of Health

Research Funding | Research at NHGRI | Health | Education | Issues in Genetics | Newsroom | Careers & Training | About | For You

NHGRI Division of Intramural Research
[Research Home Page](#)

Histone Database
[Home/Search](#)
[Histone Protein Sequences](#)
[Multiple Sequence Alignments](#)
[Human Histone Gene Complement](#)
[Non-Histone Proteins Containing the Histone Fold Motif](#)
[Structures](#)
[About the Histone Database](#)

Histone Database

Search

Search Sequence Headers for:

Sequence Fragment: PRK

Histone Type: H2B

Organism: Parechinus angulosus

Data Set: Redundant Set Only (non-archaeal)

Search

B

genome.gov
National Human Genome Research Institute
National Institutes of Health

Research Funding | Research at NHGRI | Health | Education | Issues in Genetics | Newsroom | Careers & Training | About | For You

NHGRI Division of Intramural Research
[Research Home Page](#)

Histone Database
[Home/Search](#)
[Histone Protein Sequences](#)
[Multiple Sequence Alignments](#)
[Human Histone Gene Complement](#)
[Non-Histone Proteins Containing the Histone Fold Motif](#)
[Structures](#)
[About the Histone Database](#)

Histone Database

Your search returned 2 entries.

Clicking on the protein name, the browser will display the FASTA format detail page for that particular histone protein. Columns may be resorted by clicking on any of the column headers.

To obtain a FASTA-formatted list of sequences, click on the checkboxes next to the sequences of interest, then click "Get FASTA-formatted sequence".

Get FASTA-formatted sequence

All	GI	Accession	Histone Type	Protein Name	Organism
<input type="checkbox"/>	108885304	P02291.2	H2B	H2BS2_PARAN RecName: Full=Histone H2B.2_sperm	Parechinus angulosus
<input type="checkbox"/>	108885306	P02292.2	H2B	H2BS3_PARAN RecName: Full=Histone H2B.3_sperm	Parechinus angulosus

USA.gov | Privacy | Copyright | Contact | Accessibility | Site Map | Staff Directory | FOIA

Figure 2. Histone Database query and results. The Histone Database main page displays a search engine that allows users to find histone sequences from a large variety of organisms. Additionally, users have the possibility of exploring other features to access complete collections of Histone Protein Sequences, Multiple Sequence Alignments, The Human Histone Gene Complement, Non-Histone Proteins Containing the Histone fold Motif and Histone Structures. The upper panel shown (A) presents the criteria used for the query, which requires the sequence to contain a fragment with amino acids PRK from the angulate sea urchin (*Parechinus angulosus*) histone H2B. The search results presented in the lower panel (B) include two protein sequences that meet the criteria specified by the query.

They are also available as a series of multiple sequence alignments, one for each class of proteins. A number of search engines can be used to query the database in several different ways: by protein family, organism, keyword or based on a sequence pattern (Figure 2). Each histone sequence for which three-dimensional structure data is available is linked to the corresponding entry in both PDB and the Molecular Modeling Database (MMDB) (35).

The Histone Sequence Database has been expanded significantly since its last update (23) (Table 1). However, the expansion is not proportional for each of the core histones.

The H3 sequences, which contain a large number of variants with specialized roles in chromosome segregation and transcription, show an increase over 400% since the last database update. Similarly, the H2A core histone sequences that include variants with specialized functions in DNA repair and transcription regulation show an increase over 200% since the last update. In contrast, we observe a more modest growth in sequence numbers for the relatively invariant H4 and H2B core histones.

The Histone Sequence Database now includes archaeal histone sequences. The current update contains 182

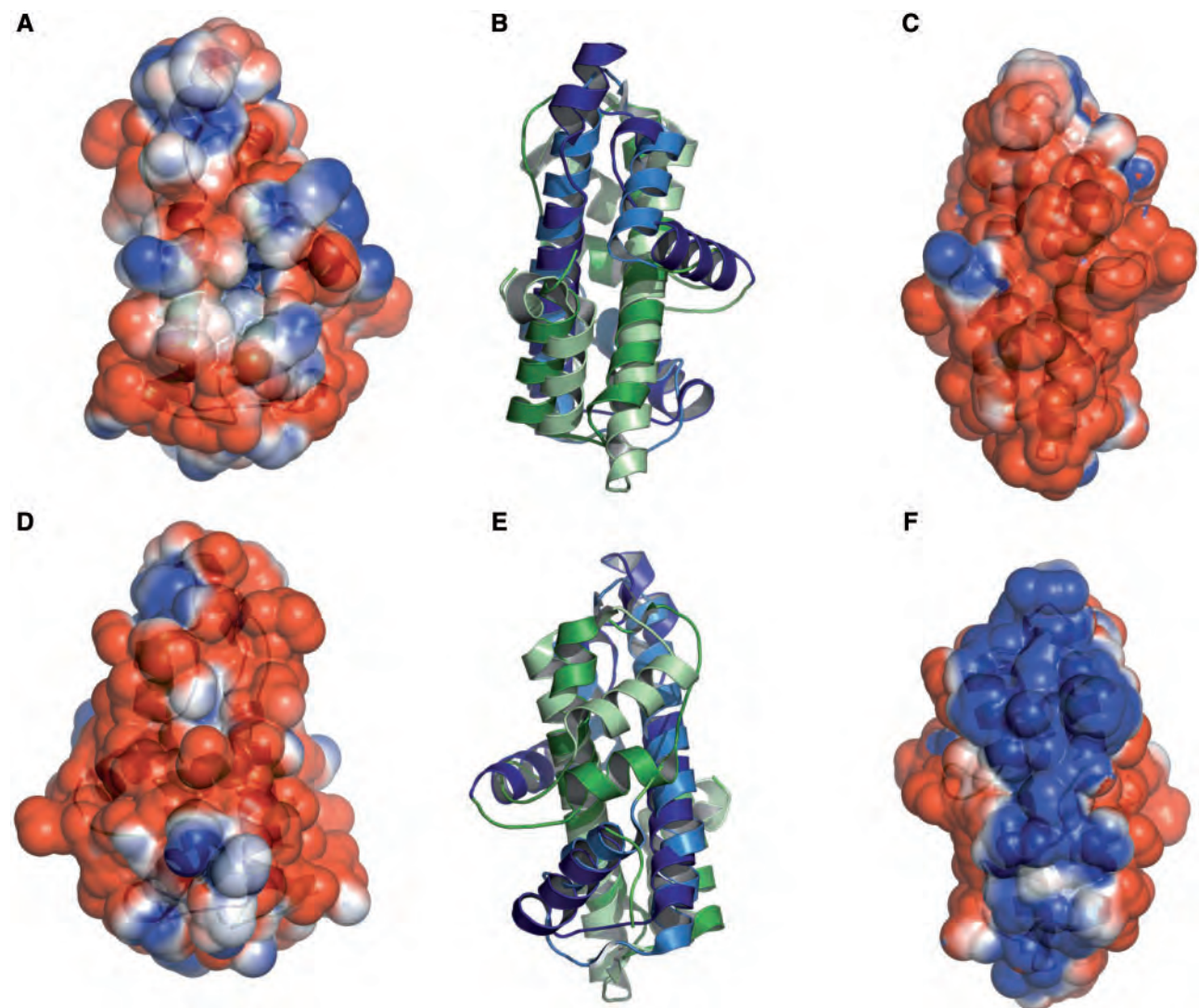


Figure 3. Histone-like folds in *A. aeolicus* and *M. kandleri*. Protein Aq_328 from the hyperthermophilic bacterium *A. aeolicus* (PDB:1R4V) and archaeal histone from *M. kandleri* (PDB:1F1E) have two histone like folds. These are colored as dark blue and dark green (for 1R4V) and light blue and light green (for 1F1E). The electrostatic surface potential ranges from $+2 \text{ kTe}^{-1}$ (blue) to -2 kTe^{-1} (red). (A) and (D) the front and back views, respectively, of the electrostatic surface potential of Protein Aq_328. (B) and (E) superimposed structures of protein Aq_328 and the archaeal histone from *M. kandleri*. (C) and (F) the front and back views, respectively, of the electrostatic surface potential of the archaeal histone from *M. kandleri*. The figures were generated with PyMOL (49) and the APBS plug-in for PyMOL (50).

sequences from 89 archaeal organisms, which includes members of all classified archaeal phyla (i.e. euryarchaeota, crenarchaeota, nanoarchaeota, korarchaeota and the newly proposed phylum thaumarchaeota). The presence of histone folds in all classified archaeal phyla indicates that the histone fold originated before the archaeal and eukaryotic lineage divergence (29). Most of the archaeal histones have a single histone fold domain; however, there are a number of sequences that contain two histone folds, with the C-terminal histone fold sharing higher sequence similarity with archaeal histones with a single histone fold. Archaeal histones containing two histone folds have been proposed as intermediates between archaeal and eukaryotic histones (36, 37), where both core histones H3 and H4 would have originated at the same time, followed by a second event that gave rise to core histones H2A and H2B. In the current release of the Histone Sequence Database, archaeal histones with two histone folds are confined to two distinct branches: Halobacteriaceae and the hyperthermophilic methanogen *Methanopyrus kandleri*. Although archaeal histones containing two histone folds have been previously identified in these lineages, it is not clear how these histones could also contribute to pack DNA in extreme high temperature or high salinity environments.

Structural comparisons confirmed the presence of the histone fold in the extreme bacterial thermophile *Aquifex aeolicus* (30). Additionally, the RIKEN Structural Genomics/Proteomics Initiative (RSGI) (38) has solved two *Thermus thermophilus* structures for a protein that also contain the histone fold (PDB:1WWI and PDB:1WWS). The histone fold was also found in diverse types of bacteria, including aquificales, ϵ -proteobacteria, thermaceae, actinobacteria and nostocaceae. This suggests that the histone fold appeared in bacteria by lateral gene transfer (29, 39). Interestingly, the structure from *T. thermophilus* (PDB:1WWS), predicted to be a dimer, is strikingly similar to the H3–H4 tetramer. However, an analysis of the electrostatic surface potential for protein Aq_328 from the hyperthermophilic bacterium *A. aeolicus* (PDB:1R4V) and archaeal histone from *M. kandleri* (PDB:1F1E) (Figure 3) reveals the DNA binding surface in the archaeal histone (Figure 3F) but shows no conservation of any of the DNA-binding residues present in both archaeal and eukaryotic histones (Figures 3A and 3D) (29). Therefore, it is possible that histone fold-like bacterial proteins have functions unrelated to DNA binding. However, it is likely that the histone-like fold is used as a dimerization domain in these species.

Conclusions

Researchers studying chromatin structure and function have traditionally relied on the Histone Sequence Database to explore the taxonomic breadth of histones

and their variants (40–43). Others have focused on epigenetics and transcriptional regulation and use the database to discover newly reported core histones and histone-fold-containing proteins (44–48). The Histone Database continues to be a comprehensive bioinformatic resource that organizes and stores histone sequences and groups them into families (that now includes archaeal histones), maintains a collection of histone fold-containing sequences, and provides information on three-dimensional structures available in PDB. In the future, we will enhance our histone fold identification pipeline with state-of-the-art sequence- and structure-based methods to continue to identify new members of this biologically critical family of proteins. We also plan to integrate functional information from other publicly available Web resources.

Acknowledgments

The Histone Database update utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Maryland. (<http://biowulf.nih.gov/>).

Funding

Funding for open access charge: The Intramural Research Programs of the National Center for Biotechnology Information, National Library of Medicine and the National Human Genome Research Institute, both at the National Institutes of Health.

Conflict of interest. None declared.

References

1. van Holde, K.E. (1988) Chromatin. Springer, New York.
2. Eickbush, T.H. and Moudrianakis, E.N. (1978) The histone core complex: an octamer assembled by two sets of protein-protein interactions. *Biochemistry*, **17**, 4955–4964.
3. Cho, R.J., Campbell, M.J., Winzeler, E.A. et al. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
4. Ausio, J. (2006) Histone variants—the structure behind the function. *Brief. Funct. Genomic Proteomic*, **5**, 228–243.
5. Elsaesser, S.J., Goldberg, A.D. and Allis, C.D. (2010) New functions for an old variant: no substitute for histone H3.3. *Curr. Opin. Genet. Dev.*, **20**, 110–117.
6. Talbert, P.B. and Henikoff, S. (2010) Histone variants—ancient wrap artists of the epigenome. *Nat. Rev. Mol. Cell Biol.*, **11**, 264–275.
7. Luebben, W.R., Sharma, N. and Nyborg, J.K. (2010) Nucleosome eviction and activated transcription require p300 acetylation of histone H3 lysine 14. *Proc. Natl Acad. Sci. USA*, **107**, 19254–19259.
8. Jin, C., Zang, C., Wei, G. et al. (2009) H3.3/H2A.Z double variant-containing nucleosomes mark ‘nucleosome-free regions’ of active promoters and other regulatory regions. *Nat. Genet.*, **41**, 941–945.
9. Kouzarides, T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.

10. Schones,D.E. and Zhao,K. (2008) Genome-wide approaches to studying chromatin modifications. *Nat. Rev. Genet.*, **9**, 179–191.
11. Arents,G., Burlingame,R.W., Wang,B.C. et al. (1991) The nucleosomal core histone octamer at 3.1 Å resolution: a tripartite protein assembly and a left-handed superhelix. *Proc. Natl Acad. Sci. USA*, **88**, 10148–10152.
12. Baxevanis,A.D., Arents,G., Moudrianakis,E.N. et al. (1995) A variety of DNA-binding and multimeric proteins contain the histone fold motif. *Nucleic Acids Res.*, **23**, 2685–2691.
13. Luger,K., Mader,A.W., Richmond,R.K. et al. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–260.
14. Davey,C.A., Sargent,D.F., Luger,K. et al. (2002) Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.*, **319**, 1097–1113.
15. Kasinsky,H.E., Lewis,J.D., Dacks,J.B. et al. (2001) Origin of H1 linker histones. *FASEB J.*, **15**, 34–42.
16. Bustin,M., Catez,F. and Lim,J.H. (2005) The dynamics of histone H1 function in chromatin. *Mol. Cell.*, **17**, 617–620.
17. Khochbin,S. (2001) Histone H1 diversity: bridging regulatory signals to linker histone function. *Gene*, **271**, 1–12.
18. Marino-Ramirez,L., Kann,M.G., Shoemaker,B.A. et al. (2005) Histone structure and nucleosome stability. *Expert Rev. Proteomics*, **2**, 719–729.
19. Marino-Ramirez,L., Jordan,I.K. and Landsman,D. (2006) Multiple independent evolutionary solutions to core histone gene regulation. *Genome Biol.*, **7**, R122.
20. Henikoff,S. and Furuyama,T. (2010) Epigenetic inheritance of centromeres. *Cold Spring Harb. Symp. Quant. Biol.*
21. Sondermann,H., Soisson,S.M., Bar-Sagi,D. et al. (2003) Tandem histone folds in the structure of the N-terminal segment of the ras activator Son of Sevenless. *Structure*, **11**, 1583–1593.
22. Berglund,L., Bjorling,E., Oksvold,P. et al. (2008) A gene-centric Human Protein Atlas for expression profiles based on antibodies. *Mol. Cell Proteomics*, **7**, 2019–2027.
23. Marino-Ramirez,L., Hsu,B., Baxevanis,A.D. et al. (2006) The Histone Database: a comprehensive resource for histones and histone fold-containing proteins. *Proteins*, **62**, 838–842.
24. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J. et al. (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.
25. Berman,H.M., Westbrook,J., Feng,Z. et al. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
26. UniProt. (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
27. Wu,C.H., Yeh,L.S., Huang,H. et al. (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
28. Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.
29. Sandman,K. and Reeve,J.N. (2006) Archaeal histones and the origin of the histone fold. *Curr. Opin. Microbiol.*, **9**, 520–525.
30. Qiu,Y., Tereshko,V., Kim,Y. et al. (2006) The crystal structure of Aq_328 from the hyperthermophilic bacteria Aquifex aeolicus shows an ancestral histone fold. *Proteins*, **62**, 8–16.
31. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
32. Altschul,S.F., Madden,T.L., Schaffer,A.A. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
33. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
34. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
35. Wang,Y., Address,K.J., Chen,J. et al. (2007) MMDB: annotating protein sequences with Entrez's 3D-structure database. *Nucleic Acids Res.*, **35**, D298–D300.
36. Fahrner,R.L., Cascio,D., Lake,J.A. et al. (2001) An ancestral nuclear protein assembly: crystal structure of the Methanopyrus kandleri histone. *Protein Sci.*, **10**, 2002–2007.
37. Malik,H.S. and Henikoff,S. (2003) Phylogenomics of the nucleosome. *Nat. Struct. Biol.*, **10**, 882–891.
38. Aoki,M., Matsuda,T., Tomo,Y. et al. (2009) Automated system for high-throughput protein production using the dialysis cell-free method. *Protein Expr. Purif.*, **68**, 128–136.
39. Alva,V., Ammelburg,M., Soding,J. et al. (2007) On the origin of the histone fold. *BMC Struct. Biol.*, **7**, 17.
40. Gonzalez-Romero,R., Rivera-Casas,C., Ausio,J. et al. (2010) Birth-and-death long-term evolution promotes histone H2B variant diversification in the male germinal cell line. *Mol. Biol. Evol.*, **27**, 1802–1812.
41. Eirin-Lopez,J.M., Gonzalez-Romero,R., Dryhurst,D. et al. (2009) The evolutionary differentiation of two histone H2A.Z variants in chordates (H2A.Z-1 and H2A.Z-2) is mediated by a stepwise mutation process that affects three amino acid residues. *BMC Evol. Biol.*, **9**, 31.
42. Potoyan,D.A. and Papoian,G.A. (2011) Energy landscape analyses of disordered histone tails reveal special organization of their conformational dynamics. *J. Am. Chem. Soc.*, **133**, 7405–7415.
43. Ozboyaci,M., Gursoy,A., Erman,B. et al. (2011) Molecular recognition of H3/H4 histone tails by the tudor domains of JMJD2A: a comparative molecular dynamics simulations study. *PLoS One*, **6**, e14765.
44. Kolarik,C., Klinger,R. and Hofmann-Apitius,M. (2009) Identification of histone modifications in biomedical text for supporting epigenomic research. *BMC Bioinformatics*, **10** (Suppl. 1), S28.
45. Sun,X.J., Xu,P.F., Zhou,T. et al. (2008) Genome-wide survey and developmental expression mapping of zebrafish SET domain-containing genes. *PLoS One*, **3**, e1499.
46. Shultz,R.W., Tatineni,V.M., Hanley-Bowdoin,L. et al. (2007) Genome-wide analysis of the core DNA replication machinery in the higher plants Arabidopsis and rice. *Plant Physiol.*, **144**, 1697–1714.
47. Weidenbach,K., Gloer,J., Ehlers,C. et al. (2008) Deletion of the archaeal histone in Methanosarcina mazei Go1 results in reduced growth and genomic transcription. *Mol. Microbiol.*, **67**, 662–671.
48. Huda,A., Marino-Ramirez,L. and Jordan,I.K. (2010) Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mob. DNA*, **1**, 2.
49. Schrödinger, (2010) The PyMOL Molecular Graphics System, Version 1.3.
50. Lerner,M.G. and Carlson,H.A. (2009) APBS Plugin for PyMOL, <http://www.pymolwiki.org/index.php/APBS>.

Development and Characterization of Microsatellite Markers for the Cape Gooseberry *Physalis peruviana*

Jaime Simbaqueba¹, Pilar Sánchez², Erika Sanchez¹, Victor Manuel Núñez Zarantes¹, Maria Isabel Chacon², Luz Stella Barrero^{1,3}, Leonardo Mariño-Ramírez^{1,3,4*}

¹ Plant Molecular Genetics Laboratory, Center of Biotechnology and Bioindustry (CBB), Colombian Corporation for Agricultural Research (CORPOICA), Bogota, Colombia,

² Facultad de Agronomía, Universidad Nacional de Colombia, Bogotá, Colombia, ³ PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia,

⁴ Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America

Abstract

Physalis peruviana, commonly known as Cape gooseberry, is an Andean *Solanaceae* fruit with high nutritional value and interesting medicinal properties. In the present study we report the development and characterization of microsatellite loci from a *P. peruviana* commercial Colombian genotype. We identified 932 imperfect and 201 perfect Simple Sequence Repeats (SSR) loci in untranslated regions (UTRs) and 304 imperfect and 83 perfect SSR loci in coding regions from the assembled *Physalis peruviana* leaf transcriptome. The UTR SSR loci were used for the development of 162 primers for amplification. The efficiency of these primers was tested via PCR in a panel of seven *P. peruviana* accessions including Colombia, Kenya and Ecuador ecotypes and one closely related species *Physalis floridana*. We obtained an amplification rate of 83% and a polymorphic rate of 22%. Here we report the first *P. peruviana* specific microsatellite set, a valuable tool for a wide variety of applications, including functional diversity, conservation and improvement of the species.

Citation: Simbaqueba J, Sánchez P, Sanchez E, Núñez Zarantes VM, Chacon MI, et al. (2011) Development and Characterization of Microsatellite Markers for the Cape Gooseberry *Physalis peruviana*. PLoS ONE 6(10): e26719. doi:10.1371/journal.pone.0026719

Editor: I. King Jordan, Georgia Institute of Technology, United States of America

Received: August 11, 2011; **Accepted:** October 3, 2011; **Published:** October 21, 2011

This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Funding: Support for this research was provided by a grant from the Colombian Ministry of Agriculture Contract Nos. 054/08072-2008L4787-3281 to LSB and 054/08190-2008L7922-3322 to VMNZ. This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and National Center for Biotechnology Information. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: marino@ncbi.nlm.nih.gov

Introduction

Physalis peruviana commonly known as Cape gooseberry or golden berry is an Andean tropical fruit from the *Solanaceae* family native to South American countries including Colombia, Ecuador and Peru. *Physalis peruviana* grows wild in various parts of the Andes, typically 2,200 meters above sea level. The Cape gooseberry was known to the Incas but their origins are not clear, after Christopher Columbus the Cape gooseberry was introduced into Africa and India [1]. In Colombia, over the last three decades, *P. peruviana* went from being a neglected species to be the most promissory and successful exotic fruit for national and international markets; thus, since 1991, the Cape gooseberry market has been growing annually and in 2007 exports brought USD 34 million into the country. The main consumers of the Colombian Cape gooseberry are Europe with 97%, along with Asia and the United States with the remaining 3% [2]. The commercial interest in this fruit has grown due to its nutritional properties related to high vitamins content, minerals and antioxidants as well as its anti-inflammatory, anti-cancer and other medicinal properties [3,4,5,6,7,8].

Despite growing interest in the Cape gooseberry, little is known about its genetic diversity and population structure. The collections kept in germplasm banks have been partially evaluated for morphologic and agronomic traits [9,10,11]. Although it has been reported that Cape gooseberry is a diploid species with

2n = 48 [12]; different chromosome numbers might exist among genotypes since 2n = 24 has been reported for wild ecotypes, 2n = 32 for the cultivated Colombia ecotype and 2n = 48 for the cultivated Kenya ecotype [13]. The genetic diversity of the Cape gooseberry at the molecular level has been poorly studied, to our knowledge there is only one report applying dominant markers RAMs (Random Amplified Microsatellites) in 43 individuals from five geographical regions in Colombia suggesting high heterozygosity and genetic diversity [14]. Additionally, in our experience, the use of heterologous microsatellite markers previously developed for several other *Solanaceae* species have not been successful in identifying polymorphic markers in Cape gooseberry.

Microsatellites or SSRs are defined as highly variable DNA sequences composed of tandem repeats of 1–6 nucleotides with co-dominant inheritance which have become the markers of choice for a variety of applications including characterization and certification of plant materials, identification of varieties with agronomic potential, genetic mapping, assistance in plant-breeding programs, among others [15,16,17,18,19]. However, no SSR markers specific for *P. peruviana* have been developed. The genetic analysis with microsatellites is simple and robust, although their identification and development present significant challenges in emerging species [16,20]. According to the origin of the sequences used for the initial identification of simple repeats, SSRs are divided in two categories: Genomic SSRs which are derived

Table 1. Plant material used for SSR development and characterization.

Species	Work Code	Accession/Common Name	Accession Code	Origin	
				Source/region	Country
<i>P. peruviana</i>	1	ILS 3804*	09U086-1	CORPOICA/Ambato	Ecuador
<i>P. peruviana</i>	2	Ecotype Kenia	09U215-1	Universidad de Nariño/*NA	Colombia
<i>P. peruviana</i>	3	Ecotype Colombia	09U216-1	Universidad de Nariño/NA	Colombia
<i>P. floridana</i>	4	ILS 1437*	09U139-1	Botanical Garden of Birmingham/NA	U.K.
<i>P. peruviana</i>	5	Novacampo (commercial)	09U 274-1	CORPOICA/Cundinamarca	Colombia
<i>P. peruviana</i>	6	ILS 3807*	09U089-1	CORPOICA/Antioquia	Colombia
<i>P. peruviana</i>	7	ILS 3826*	09U108-1	CORPOICA/Antioquia	Colombia
<i>P. peruviana</i>	8	ILS 3817*	09U099-1	CORPOICA/Caldas	Colombia

ILS* = Introduction maintained at La Selva Research Center, CORPOICA; NA = Not available; *NA = Not available (*in vitro* propagated material).
doi:10.1371/journal.pone.0026719.t001

from random genomic sequences and EST-SSRs derived from expressed sequence tags or from coding sequences. Genomic SSRs are not expected to have neither genic function nor close linkage to transcriptional regions, while EST-SSRs and coding-SSRs are tightly linked with functional genes that may influence certain important agronomic characters. The *de novo* identification of simple sequence repeats has usually involved large-scale sequencing of genomic, SSR-enriched genomic or EST libraries, which are expensive, laborious and time-consuming. Next generation sequencing technologies have enabled rapid identification of SSR loci derived from ESTs which can be identified in any emergent species [17,19,21].

The goal of the present study was to identify polymorphic SSR loci using the assembled leaf transcriptome sequences from a commercial Colombian ecotype of *P. peruviana* developed in our laboratory (<http://www.ncbi.nlm.nih.gov/bioproject/67621>). Imperfect as well as perfect repeat searches in non-coding or untranslated regions (UTRs) were performed. From these loci, primers were designed for amplification of UTR SSR loci. The effectiveness of these primers was tested via PCR in seven *P. peruviana* accessions, among them, the ecotypes Colombia, Kenya and Ecuador, as well as one closely related species *Physalis floridana*. The molecular markers developed here are valuable tools for assessing functional diversity, aid in species conservation and plant breeding programs.

Materials and Methods

SSR loci identification and marker development

A collection of *Physalis peruviana* leaf transcript sequences was used as the source for SSR development (Transcriptome Shotgun Assembly (TSA) Database, GenBank Accession numbers JO124085-JO157957). The transcripts were compared for sequence similarity with the non-redundant protein sequences database from NCBI using BLASTX. SSR loci were searched in both coding and non-coding sequences. Candidate SSR loci were identified using Phobos [22] in both coding and non-coding sequences using perfect and imperfect repeat searches with a minimum length of 18 bp for dinucleotides, 24 bp for tri and tetranucleotides, 30 bp for pentanucleotides and 36 bp for hexanucleotide repeats.

Primer design and amplification of SSR loci by PCR

Primer3 version 0.4.0 [23] was used to design primers for microsatellite amplification in *P. peruviana*. In addition, the

oligocalculator - SIGMA Aldrich (<http://www.sigma-genosys.com/calc/DNACalc.asp>) was used to predict secondary structures (i.e. hairpins, primer dimers) for each primer pair designed. To determine the success of the microsatellite primer design, we carried out PCR tests to amplify the SSR loci in seven *P. peruviana* accessions (including Kenya, Ecuador and Colombia ecotypes) and one *Physalis floridana* accession, a closely related species (Table 1). The following PCR conditions were used: 1X PCR buffer: 1.5 to 3 mM MgCl₂ depending on the primer pair, 0.2 μM dNTPs, 0.2 to 0.3 μM of each primer (depending on the primer pair), 0.05 U/μl *Taq* polymerase and 25 ng of genomic DNA, in a 15 μl reaction volume. The temperature conditions were 95°C for 3 minutes followed by 35 cycles of 95°C for 30 seconds, 50 to 52°C (depending on the primer pair) for 30 seconds and 72°C for 90 seconds, and a final extension of 72°C for 8 minutes. The PCR amplification products were analyzed by polyacrylamide gel electrophoresis (PAGE).

Gene Ontology analysis of SSR loci

A gene ontology (GO) analysis was performed using blast2go [24] with the assembled transcript sequences containing the 30 polymorphic SSRs described here. These sequences were compared with the UniProtKB/Swiss-Prot database with a cutoff e-value of 1×10^{-5} .

Table 2. SSR loci identified in *Physalis peruviana* leaf Expressed Sequence Tags (ESTs).

Repeat Type	Perfect			Imperfect			Frequency	
	CDS	UTRs	Total	CDS	UTRs	Total		
Dinucleotide	-	34	34	2	98	100	134	8%
Trinucleotide	36	81	117	178	249	427	544	36%
Tetranucleotide	1	16	17	13	69	82	99	7%
Pentanucleotide	-	6	6	47	160	207	213	14%
Hexanucleotide	46	64	110	64	356	420	530	35%
Total	83	201	284	304	932	1236	1520	-
Frequency	6%	13%	19%	20%	61%	81%		

The number of SSR loci identified at coding sequences (CDS) and Untranslated Regions (UTRs) by using perfect and imperfect repeat search criteria.
doi:10.1371/journal.pone.0026719.t002

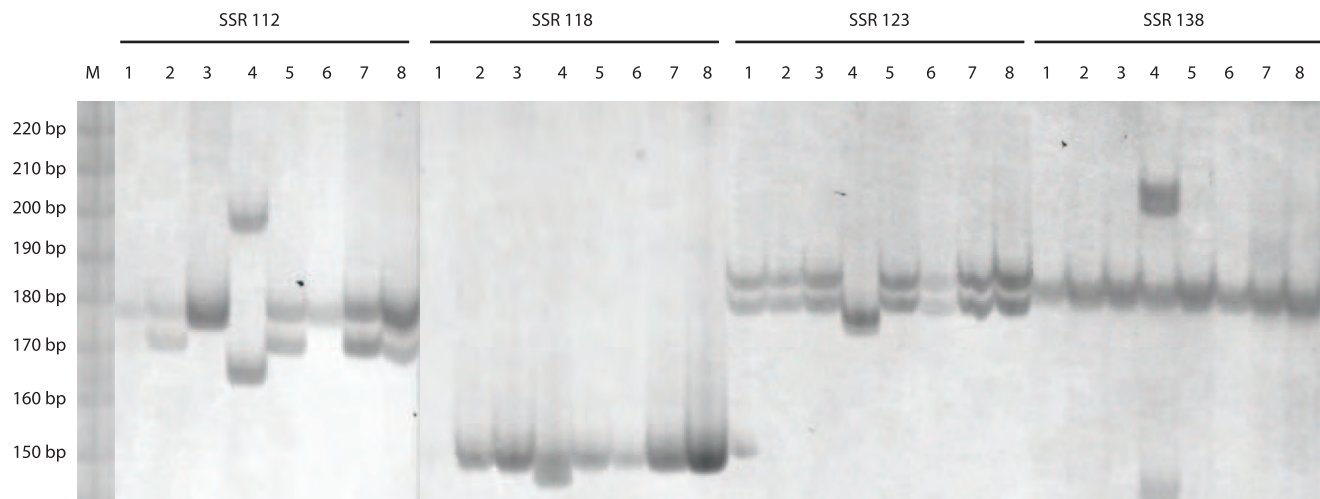


Figure 1. SSR alleles in eight *Physalis* genotypes and four polymorphic loci. The polymorphic SSR loci were visualized in 6% polyacrylamide gels, samples 1–8 correspond to the work code shown in Table 1. M= Molecular size marker, 10 bp DNA Ladder (Invitrogen, Carlsbad, CA). doi:10.1371/journal.pone.0026719.g001

Results

Identification of SSR loci in *P. peruviana*

A total of 1,520 SSR loci were identified and a large fraction were located in UTRs (74%) as compared to coding sequences (CDS) with 26%. The highest number of SSR loci found contained trinucleotide and hexanucleotide repeats with 544 (36%) and 530 (35%) respectively (Table 2).

Microsatellite primer design and PCR analysis

The SSR loci selected for primer design were located at UTRs and identified with an imperfect repeat search to increase the probabilities for finding polymorphisms within the individuals analyzed. Using this strategy a total of 162 primers pairs were designed. A successful PCR amplification was obtained for 138 (83%) of the 162 primers designed from microsatellite loci using seven *P. peruviana* and one *P. floridana* genotype (Table 1). Polymorphisms among the eight genotypes were observed for 30 (22%) loci whereas the remaining 108 loci were monomorphic (Figure 1, Tables 3 and 4).

Functional relationships of polymorphic SSR markers

A significant GO annotation was found for 10 of the 30 markers, which are related to 43 different ontology terms, of these 27 (67%) were related to biological process, 11 (25%) to molecular function and 5 (8%) to cellular component (Table 5).

Discussion

Here we present the first collection of EST-derived microsatellite markers in *Physalis peruviana*. The highest number of SSR loci found contained trinucleotide and hexanucleotide repeats (Table 2), which is consistent with results reported in Solanaceae and other plant species [19,20,25,26,27,28,29,30,31]. 1,236 out of 1,520 SSR loci are composed of imperfect repeats increasing the probability of polymorphism among *Physalis* species. This inference is bolstered by the fact that 30 of the 162 imperfect SSRs (22%) were polymorphic in the panel of 8 accessions from *P. peruviana* and the related species *P. floridana* (Table 1), suggesting the potential utility of these genetic based SSR markers for future studies. i.e. germplasm diversity and breeding applications [17,19,32].

Our results show that most of the SSR loci were located at UTRs (Table 2) in agreement with the results reported by Morgante and others [27] who hypothesize that in plants most of the SSR loci from transcribed regions are distributed along the UTRs. Increased numbers of SSR loci at UTRs could be related to changes in transcription (5'UTRs) or RNA silencing (3'UTRs), which are sources of variation among species [18,19,20,29,30]. Cereal species appear to have a different SSR distribution; Yu and others [33] found that most of the 444 EST derived SSR markers (62%) were located at coding regions, while 38% were located at UTRs.

Since the SSR loci found in this study were derived from genes, they may be related to some traits of interest [18,20,27] such as resistance to *Fusarium oxysporum*, which is one of the main constraints for Cape gooseberry production at the commercial level. According to the functional annotation obtained by the GO analysis, two polymorphic SSR markers (SSR54 and SSR77 respectively) were related with proteins involved in defense responses to pathogens such as programmed cell death and ethylene as well as jasmonic acid pathways. These two polymorphic SSR makers would be useful in *P. peruviana* breeding programs focused on *F. oxysporum* resistance.

The high rate of successful PCR amplification for the primer pairs designed (84%, Table 4) is related to the fact that these loci are specific to *P. peruviana* and they were also developed from genes, increasing the transferability within species of the same genus i.e. *P. floridana*. These results are in agreement with Zeng

Table 3. Polymorphisms in *Physalis peruviana* SSR loci.

SSR Type	Polymorphic	Monomorphic	Total
Dinucleotide	19	53	72
Trinucleotide	10	39	49
Tetranucleotide	-	5	5
Pentanucleotide	1	1	2
Hexanucleotide	-	10	10
Total	30	108	138

doi:10.1371/journal.pone.0026719.t003

Table 4. Allelic variation in 30 *Physalis peruviana* SSR loci.

Polymorphic loci	Forward primer (5'-3')	Reverse primer (5'-3')	PCR conditions		Alleles (pb)		Repeat type	Location
			Primer [μM]	MgCl ₂ [mM]	°Tm	Expected size	Range size observed	
SSR1	AGAGACTCATTGTTGCT	TGAGGTGTGGATGTTTCT	0,2	2	50	206	170 210	AT 3' UTR
SSR2	CATTGGTTTCGATCCAT	AGACAAGCTAGGGAAAGG	0,2	2	50	237	230 250	AG 3' UTR
SSR9	TGCTCGAGTTTTCAGGTTTC	GCAGTGGTAAAGTTGAGAGACG	0,2	2	50	193	220 240	AG 5' UTR
SSR10	GCTTCCTATTGTTGCTGA	ACTTTGGGTTTCGGGAATTG	0,2	2	50	185	170 190	AT 3' UTR
SSR11	CAGCTGAATAAGAGAGTGATTGG	CCCTCTTTTCTCTCCGAGT	0,2	2	50	180	180 210	AG 3' UTR
SSR13	GCGGAATCCATTGTTTTC	CCGATGAGATATAGTCACGCAAA	0,2	2	50	190	160 210	AC 5' UTR
SSR14	TGAACCCATCTAGCTGAACG	TGGTGTGTTCTTACAATCCAT	0,2	1,5	50	204	200 220	AT 3' UTR
SSR15	GCTTGTGATCAGCTTCTTTC	TGGATCATAACCTTGCTAATGC	0,2	1,5	50	172	160 180	AT 3' UTR
SSR18	CAGAGTGAATACCTTGACGAA	TGTCCATTTTGTGCGCAAT	0,2	1,5	50	179	180 230	AC 3' UTR
SSR20	GCACATCACATAAAGTATCTTCTCA	TGCTCGTGTGTTCTGCTATG	0,2	1,5	50	270	170 220	AT 3' UTR
SSR36	ATGAACACATGTCGGAGG	GGGGATCCAAACGAAGTGA	0,2	1,5	52	211	170 240	AG 3' UTR
SSR37	CCAAGTGAATCAACACACAGC	CCACACTGAAAGGAGGATCTG	0,3	2	50	212	260 330	AG 3' UTR
SSR54	CGGTGGTATGCTTACAAAGAT	GCATTCCTACTGTTTTAACTTCC	0,2	1,5	50	197	190 210	AC 3' UTR
SSR55	CACCTACATAGGCAGCCAAA	ATTGTGGCGGAGGAG	0,2	1,5	50	183	200 210	AG 5' UTR
SSR57	AGTGAAGACAGCCATTCT	GGCGAAGCTGAATTGAAAA	0,2	1,5	50	183	200 210	AT 3' UTR
SSR67	GCTCTGTTCATTATTCACA	GCAGTGTGGGATCAATCAAT	0,2	1,5	50	207	180 240	AG 3' UTR
SSR68	GAAGCAACAACACTACACCCAAA	AAGCTCGGATTCATAGCA	0,2	1,5	50	187	160 220	AG 3' UTR
SSR72	GTGTCGCGAGTTCTTCAAA	CCGCCGTTACTCTTAATCA	0,2	1,5	50	158	130 170	AG 3' UTR
SSR77	CATACCATAACTCCCATCTCTC	TGCCGATTCTGATTCTTCC	0,2	1,5	50	216	170 200	AT 5' UTR
SSR92	TGTTTTGAGGATCAAGAAAGAA	GTGGTATCAACGCAAGAGTGG	0,25	2,5	50	205	180 210	AAG 3' UTR
SSR107	CATCCAACACCAGAAATACGC	TCCAACCTTATCAATTTCTCCAC	0,2	1,5	50	206	220 250	AAG 5' UTR
SSR110	CACCCATATCCCAATCTCTTC	GGGTAATTTTACGGGGAAT	0,2	1,5	50	198	170 200	CTT 3' UTR
SSR112	CTAGCTACCACTTGCACA	CAGTGAAGCCCTCAAGATCC	0,2	1,5	50	203	200 220	TCT 3' UTR
SSR118	AATCAAGGTCAGAAAGAAATGG	GCAAGAATGGATGGGTGT	0,2	1,5	50	180	130 180	AAG 5' UTR
SSR121	AGCAACCTCCCAATCAGCTA	TGGTGAGTAAATGGGGAAA	0,2	1,5	50	189	170 190	ATC 3' UTR
SSR123	TCACTGGAGCGGTATATCT	GCGATCTCACCACCTCTC	0,2	1,5	50	216	190 210	ATC 5' UTR
SSR126	TCCAAAAGAAAACAAAACACT	TTGAATGCATGTTTATGGA	0,2	1,5	50	202	190 200	AGC 5' UTR
SSR127	TTGGTTGGCATACTGCAA	GGTTTGAACCTCTCATGCTG	0,2	1,5	50	180	140 160	AAT 5' UTR
SSR138	TCCGATCACTACTCAGCACG	CAATCGGTTGTGAATCGGGT	0,2	1,5	50	138	130 160	AAT 3' UTR
SSR146	AGGCTAATGAGGACGAAGCA	GGTTGCATTACAAAGCACTGA	0,2	1,5	50	187	160 210	AAAAG 3' UTR

doi:10.1371/journal.pone.0026719.t004



Table 5. Functional annotation of 10 *P. peruviana* contigs containing polymorphic SSR markers.

SSR Marker	GO Category: ID	Functional Annotation
SSR2	P:0006350	Transcription
SSR37	F:0016301	Kinase activity
	C:0005886	Plasma membrane
SSR54	P:0006952	Defense response
	P:0012501	Programmed cell death
	C:0044464	Cell part
	F:0000166	Nucleotide binding
SSR55	P:0051865	Protein autoubiquitination
	F:0004842	Ubiquitin-protein ligase activity
	P:0048437	Floral organ development
	P:0046621	Negative regulation of organ growth
SSR77	P:0009789	Positive regulation of abscisic acid mediated signaling pathway
	P:0006979	Response to oxidative stress
	P:0052544	Callose deposition in cell wall during defense response
	P:0009753	Response to jasmonic acid stimulus
	P:0031348	Negative regulation of defense response
	P:0008219	Cell death
	P:0009651	Response to salt stress
	P:0042742	Defense response to bacterium
	P:0009926	Auxin polar transport
	P:0010119	Regulation of stomatal movement
	P:0009408	Response to heat
	F:0005515	Protein binding
	P:0010150	Leaf senescence
	P:0048765	Root hair cell differentiation
	P:0009871	Jasmonic acid and ethylene-dependent systemic resistance, ethylene mediated signaling pathway
	P:0001736	Establishment of planar polarity
	P:0050832	Defense response to fungus
	P:0010182	Sugar mediated signaling pathway
SSR92	F:0004674	Protein serine/threonine kinase activity
	P:0045449	Regulation of transcription
	P:0007169	Transmembrane receptor protein tyrosine kinase signaling pathway
	F:0005524	ATP binding
	F:0003700	Transcription factor activity
	P:0010030	Positive regulation of seed germination
	P:0006468	Protein amino acid phosphorylation
SSR110	C:0044444	Cytoplasmic part
SSR126	F:0005488	Binding
	F:0003824	Catalytic activity
SSR138	F:0016740	Transferase activity
SSR146	C:0005730	Nucleolus
	C:0016020	Membrane
	F:0003677	DNA binding

Gene ontology (GO) functional Categories: **C** = Cellular component, **F** = Molecular function, **P** = Biological process.
doi:10.1371/journal.pone.0026719.t005

et al. and Csencsics *et al.* [19,21], who used full-length cDNA and ESTs and found rates of successful PCR amplification larger than 80%.

This study reports the first set of microsatellite markers developed for *P. peruviana* and related species. A total of 1,520 SSR loci were identified, including 932 imperfect SSRs located at

UTRs. From these loci a total of 162 SSR primers were developed to assay their utility as microsatellite markers in a panel of seven accessions of *P. peruviana* and one accession of *P. floridana* by PCR amplification. A total of 138 (83%) primer markers amplified, with a polymorphism rate of 22%. The markers developed here can be used in plant breeding programs that may ultimately lead to superior phenotypic characteristics such as increase in fruit size, reduction in the tendency to split during transport, reduction in the plant susceptibility to pests and diseases, and improvement of fruit quality.

References

1. Popenoe H, King S, Leon J, Kalinowski L (1990) Goldenberry (cape gooseberry). Lost Crops of The Incas: Little-known Plants of the Andes with Promise for Worldwide Cultivation, ed by National Research Council National Academy Press, Washington DC. pp 241–252.
2. Bonilla MH, Arias PA, Landínez LM, Moreno JM, Cardozo F, et al. (2009) Agenda prospectiva de investigación y desarrollo tecnológico para la cadena productiva de la uchuva en fresco para exportación en Colombia- Ministerio De Agricultura y Desarrollo Rural; Proyecto Transición De La Agricultura; Universidad Nacional De Colombia CCDIAC, editor. BOGOTÁ D.C., Colombia: Ministerio de Agricultura y Desarrollo Rural.
3. Yen CY, Chiu CC, Chang FR, Chen JY, Hwang CC, et al. (2010) 4beta-Hydroxywithanolide E from *Physalis peruviana* (golden berry) inhibits growth of human lung cancer cells through DNA damage, apoptosis and G2/M arrest. *BMC Cancer* 10: 46.
4. Wu SJ, Chang SP, Lin DL, Wang SS, Hou FF, et al. (2009) Supercritical carbon dioxide extract of *Physalis peruviana* induced cell cycle arrest and apoptosis in human lung cancer H661 cells. *Food Chem Toxicol*.
5. Pinto Mda S, Ranilla LG, Apostolidis E, Lajolo FM, Genovese MI, et al. (2009) Evaluation of antihyperglycemia and antihypertension potential of native Peruvian fruits using in vitro models. *J Med Food* 12: 278–291.
6. Ramadan MF, Morsel JT (2003) Oil goldenberry (*Physalis peruviana* L.). *J Agric Food Chem* 51: 969–974.
7. Franco LA, Matiz GE, Calle J, Pinzon R, Ospina LF (2007) [Antiinflammatory activity of extracts and fractions obtained from *Physalis peruviana* L. calyces]. *Biomedica: revista del Instituto Nacional de Salud* 27: 110–115.
8. Martínez W, Ospina LF, Granados D, Delgado G (2010) In vitro studies on the relationship between the anti-inflammatory activity of *Physalis peruviana* extracts and the phagocytic process. *Immunopharmacol Immunotoxicol* 32: 63–73.
9. Lagos Burbano TC, Criollo Escobar H, Ibarra A, Hejeile H (2003) Caracterización morfológica de la colección Nariño de uvilla o uchuva *Physalis peruviana* L. *Fitotecnia Colombiana* 3: 1–9.
10. Ligarreto GA, Lobo M., Correa A (2005) Recursos genéticos del género *Physalis* en Colombia. In: Fischer G, Miranda, D, Piedrahita, W, Romero, J, eds. *Avances en cultivo, poscosecha y exportación de la uchuva (Physalis peruviana L.) en Colombia*. Universidad Nacional de Colombia (Sede Bogotá) Facultad de Agronomía ed. Bogotá: Unibiblos, Universidad Nacional de Colombia. pp 9–27.
11. Trillos González O, Cotes Torres JM, Medina Cano CI, Lobo Arias M, Navas Arboleda AA (2008) Caracterización morfológica de cuarenta y seis accesiones de uchuva (*Physalis peruviana* L.), en Antioquia (Colombia). *Revista Brasileira de Fruticultura* 30: 708–715.
12. Menzel MY (1951) The cytotaxonomy and genetics of *Physalis*. *Proceedings of the American Philosophical Society* 95: 132–183.
13. Nohra C, Rodriguez C, Bueno A (2006) Study of the cytogenetic diversity of *Physalis peruviana* L.(Solanaceae). *Acta biol Colomb* 11: 75–85.
14. Muñoz Flórez JE, Morillo Coronado AC, Morillo Coronado Y (2008) Random amplified microsatellites (RAMS) in plant genetic diversity studies. *Acta Agron* 57: 219–226.
15. Goldstein DB, Schlötterer C (1999) Microsatellites: evolution and applications (POD).
16. Scott KD, Egger P, Seaton G, Rossetto M, Ablett EM, et al. (2000) Analysis of SSRs derived from grape ESTs. *TAG Theoretical and Applied Genetics* 100: 723–726.
17. Bozhko M, Riegel R, Schubert R, Muller-Starck G (2003) A cyclophilin gene marker confirming geographical differentiation of Norway spruce populations and indicating viability response on excess soil-born salinity. *Mol Ecol* 12: 3147–3155.
18. Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends Biotechnol* 23: 48–55.
19. Zeng S, Xiao G, Guo J, Fei Z, Xu Y, et al. (2010) Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics* 11: 94.
20. Fujimori S, Washio T, Higo K, Ohtomo Y, Murakami K, et al. (2003) A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. *FEBS Lett* 554: 17–22.
21. Csencsics D, Brodbeck S, Holderegger R (2010) Cost-effective, species-specific microsatellite development for the endangered Dwarf Bulrush (*Typha minima*) using next-generation sequencing technology. *J Hered* 101: 789–793.
22. Mayer C (2008) Phobos, a Tandem Repeat Search Tool for Complete Genomes. Version.
23. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132: 365–386.
24. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674–3676.
25. Lu FH, Cho MC, Park YJ (2011) Transcriptome profiling and molecular marker discovery in red pepper, *Capsicum annuum* L. TF68. *Mol Biol Rep*.
26. Barchi L, Lanteri S, Portis E, Acquadro A, Vale G, et al. (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics* 12: 304.
27. Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* 30: 194–200.
28. Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell Mol Biol Lett* 7: 537–546.
29. Eujayl I, Sledge MK, Wang L, May GD, Chekhovskiy K, et al. (2004) *Medicago truncatula* EST-SSRs reveal cross-species genetic markers for *Medicago* spp. *Theor Appl Genet* 108: 414–422.
30. La Rota M, Kantety RV, Yu JK, Sorrells ME (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* 6: 23.
31. Luro FL, Costantino G, Terol J, Argout X, Allario T, et al. (2008) Transferability of the EST-SSRs developed on Nules clementine (*Citrus clementina* Hort ex Tan) to other Citrus species and their effectiveness for genetic mapping. *BMC Genomics* 9: 287.
32. Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 10: 967–981.
33. Yu JK, La Rota M, Kantety RV, Sorrells ME (2004) EST derived SSR markers for comparative mapping in wheat and rice. *Mol Genet Genomics* 271: 742–751.

Author Contributions

Conceived and designed the experiments: LM-R LSB MIC VMNZ. Performed the experiments: JS ES PS. Analyzed the data: LM-R LSB MIC JS ES PS. Contributed reagents/materials/analysis tools: LM-R LSB MIC VMNZ. Wrote the paper: LM-R LSB JS. Conceived the SSR in silico component: LM-R. Conceived the SSR in silico component: LM-R. Conceived and oversaw the project and its components: LSB. Conceived the SSR analyses in the laboratory: MIC. Contributed to the original concept of the project: LSB LM-R MIC VMNZ. Jointly performed the in silico identification of SSRs: JS ES. Performed SSR analyses in laboratory: PS. Jointly prepared the manuscript: JS LSB MIC LM-R. Selected plant material for analyses: MIC LSB.

Prediction of Transposable Element Derived Enhancers Using Chromatin Modification Profiles

Ahsan Huda^{1,9}, Eishita Tyagi^{1,9}, Leonardo Mariño-Ramírez^{2,3}, Nathan J. Bowen^{1,4}, Daudi Jjingo¹, I. King Jordan^{1,3*}

1 School of Biology, Georgia Institute of Technology, Atlanta, Georgia, United States of America, **2** National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, **3** PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia, **4** Ovarian Cancer Institute, Georgia Institute of Technology, Atlanta, Georgia, United States of America

Abstract

Experimentally characterized enhancer regions have previously been shown to display specific patterns of enrichment for several different histone modifications. We modelled these enhancer chromatin profiles in the human genome and used them to guide the search for novel enhancers derived from transposable element (TE) sequences. To do this, a computational approach was taken to analyze the genome-wide histone modification landscape characterized by the ENCODE project in two human hematopoietic cell types, GM12878 and K562. We predicted the locations of 2,107 and 1,448 TE-derived enhancers in the GM12878 and K562 cell lines respectively. A vast majority of these putative enhancers are unique to each cell line; only 3.5% of the TE-derived enhancers are shared between the two. We evaluated the functional effect of TE-derived enhancers by associating them with the cell-type specific expression of nearby genes, and found that the number of TE-derived enhancers is strongly positively correlated with the expression of nearby genes in each cell line. Furthermore, genes that are differentially expressed between the two cell lines also possess a divergent number of TE-derived enhancers in their vicinity. As such, genes that are up-regulated in the GM12878 cell line and down-regulated in K562 have significantly more TE-derived enhancers in their vicinity in the GM12878 cell line and vice versa. These data indicate that human TE-derived sequences are likely to be involved in regulating cell-type specific gene expression on a broad scale and suggest that the enhancer activity of TE-derived sequences is mediated by epigenetic regulatory mechanisms.

Citation: Huda A, Tyagi E, Mariño-Ramírez L, Bowen NJ, Jjingo D, et al. (2011) Prediction of Transposable Element Derived Enhancers Using Chromatin Modification Profiles. PLoS ONE 6(11): e27513. doi:10.1371/journal.pone.0027513

Editor: Yamini Dalal, National Cancer Institute, United States of America

Received: June 28, 2011; **Accepted:** October 18, 2011; **Published:** November 7, 2011

Copyright: © 2011 Huda et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: IKJ and AH were supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839) (<http://www.sloan.org/>). Eishita Tyagi was supported by the Bioinformatics program at the Georgia Institute of Technology (<http://www.bioinformatics.biology.gatech.edu/>). NJB was supported by the Integrated Cancer Research Center at the Georgia Institute of Technology (<http://ovariancancerinstitute.org/>). This research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI (<http://www.ncbi.nlm.nih.gov/>). LMR is supported by Corporación Colombiana de Investigación Agropecuaria CORPOICA (<http://www.corpoica.org.co/SitioWeb/>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: kingjordan@biology.gatech.edu

9 These authors contributed equally to this work.

Introduction

Transposable elements (TEs) are repetitive genetic sequences that can move from one location in the genome to another. TE-derived sequences are abundant in eukaryotes and make up substantial fractions of their genomic DNA. TEs have long been dismissed as selfish DNA elements that make little or no contribution to the function of their host genomes [1,2]. This idea was supported by theoretical demonstrations that TEs can persist and proliferate in a genome without providing any function or benefit to the host [3]. In the last couple of decades however, a number of anecdotal cases of TEs contributing regulatory or coding sequences to the host genome were reported. This has led to the development of a more nuanced view of TE sequences, whereby the relationship between TEs and the host genome can be characterized as a continuum ranging from extreme parasitism to obligate mutualism with their host [4,5]. Indeed, TEs have been implicated in numerous functions that benefit the human genome.

One way in which TEs can provide functional utility to the host genome is by donating enhancer sequences that can regulate the expression of host genes.

Enhancers are distal regulatory sequences, found outside of proximal promoter regions, which can increase the expression of genes by interacting with transcription factors. There are a handful of studies that provide experimental evidence for the exaptation of TE sequences as functional enhancers in the human genome. The first example comes from a study in 1993 by Hambor *et al.* which shows that an Alu element serves as part of an enhancer that up-regulates the CD8 alpha gene in accordance with its role in differentiation along the hematopoietic lymphoid lineage [6]. A few years later another study reported that an L1 element sequence donates an enhancer to up-regulate the expression of the APOC (Apolipoprotein) gene by more than 10-fold in cultured hepatocyte cells [7]. Similarly, ancient SINE elements have been shown to serve as enhancers in mammalian specific brain formation. Santangelo *et al.* demonstrated the

selection of a MAR1 element as an enhancer for the POMC (Proopiomelanocortin) gene expressed in the pituitary gland of jawed vertebrates [8]. Another gene FGF8 (fibroblast growth factor 8) has also been shown to be regulated by the AmnSINE1 element in mammalian neuronal tissues [9]. A final study by Bejerano et al. showed that an ancient SINE element drives the expression of ISL1 (insulin gene enhancer protein) in an *in-vivo* mouse enhancer assay [10].

In addition to the experimental evidence showing that individual TE sequences provide functional enhancers to host genomes, we previously found evidence to suggest that human TEs may provide numerous enhancer sequences genome-wide. Our prior analysis showed that TE sequences reside in a substantial fraction of DNaseI hypersensitive (DHS) sites [11]. The location of DHS sites signal 'open chromatin' regions which are involved in the regulation of transcription such as promoters and enhancers [12]. The genome-wide analysis of DHS revealed that 23% of these sites contain TE sequences and are associated with higher expression levels of nearby genes in CD4⁺ T-cells [11]. These data suggested that TEs may provide a large number of regulatory sequences that can increase the expression of genes in various tissues. Given the evidence from the experimental cases of TE-derived enhancers and the presence of TE sequences in DHS sites genome-wide, our goal in this study was to further explore the contribution of TEs in donating enhancers to various human cell types.

Experimentally characterized active enhancers display a distinct pattern of chromatin modifications that is significantly different from other regulatory regions as well as the genomic background [13,14,15]. Specifically, functionally active enhancers are enriched for a suite of individual histone modifications – H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac – and their enrichment patterns can be used to predict novel enhancers [13,14]. We used the chromatin signature of active enhancers to guide the search for putative TE-derived enhancers in two human hematopoietic cell lines, GM12878 and K562, characterized as part of the ENCODE project [16,17]. We employed a computational approach to identify novel enhancers by building a training set based on ChIP-Seq tag counts of the five enhancer-characteristic histone modifications found over a set of previously defined enhancer regions. Using genome-wide histone modification maps for the GM12878 and K562 cell lines, we identified hundreds of enhancers donated by TEs in each cell line. We also investigated the functional effect of these enhancers on gene expression and observed that TE-derived enhancers play a role in regulating gene expression in a cell type specific manner.

Results and Discussion

Specific chromatin modification profiles have been shown to mark functionally active enhancer regions in the human genome [13,14,15]. We employed a computational approach that uses the patterns of histone modifications to predict novel active enhancers in two human cell lines. The ENCODE project recently characterized genome-wide locations for several histone modifications in different human cell lines [16,17]. We chose two cell lines derived from the hematopoietic stem cell lineage: GM12878 and K562. GM12878 is a lymphoblastoid cell line derived from a female donor of northern and western European descent, whereas K562 is an immortalized cancer cell line derived from a northern European female patient suffering from immortalized Chronic Myelogenous Leukemia (CML). In each cell line, we analyzed the distribution of eight histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3,

H4K20me1) characterized using chromatin immunoprecipitation followed by sequencing (ChIP-Seq). Functional enhancers have also been associated with 'open chromatin' as described by DHS sites. Therefore, we also incorporated data of the genomic locations of DHS characterized by the ENCODE project in GM12878 and K562 cells [16,17].

Enhancer training set

Functionally active enhancers are marked by an enrichment of the transcription co-activator protein p300 [18,19]. As an integral part of the enhancer-associated protein complex, p300 has been found at enhancer locations across the human genome [20,21,22]. A set of p300 bound genomic locations has recently been characterized using the ChIP-chip technique in human K562 cells, and these p300 binding sites have been taken to represent a genome-wide map of functional enhancers [13,14]. In order to determine the chromatin modification profile of enhancers, we evaluated eight histone modifications at experimentally characterized p300 binding sites in the K562 cell line. We found that five of these modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac) display distinct patterns at p300 binding sites that can be used to predict putative enhancers (Figure 1). Therefore, we selected 137 of the p300 binding sites that are significantly enriched for all five modifications above the genomic background to build an enhancer training set. The training set consists of five vectors, each representing ChIP-Seq tag counts of individual histone modifications over a 10 kb region divided in 100 bp bins and summed over the 137 p300 binding sites (see Methods).

We employed two controls to validate the discriminatory power of the chromatin modification profile captured by our enhancer training set. As a first control, we compared the genomic location profiles for mapped ChIP-Seq tags of the five enhancer enriched histone modifications with the three remaining modifications around the 137 p300 binding sites. The five enhancer enriched histone modifications present in our training set (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3) display unique patterns of enrichment, with tag count peaks centered around the p300 binding sites, whereas the other three histone modifications (H3K27me3, H3K36me3, H4K20me1) do not show any specific pattern of enrichment over enhancer regions (Figure S1). As a second control, we sampled 137 random genomic sequences and compared the profiles of histone modification tag counts against those in our training set derived from p300 binding sites. We observed that random genomic locations do not display any pattern of histone modification enrichment characteristic of experimentally characterized enhancers (Figure S2). Taken together, these controls demonstrate that experimentally characterized enhancers display unique patterns of enrichment of five histone modifications, which are significantly different from the genomic background. Thus, the epigenetic histone modification profile captured by our enhancer training set possesses the discriminating features necessary to search for novel enhancers.

We attempted to further ascertain the discriminating power of our enhancer training set by performing cross-validation together with receiver operating characteristic (ROC) analysis on the genomic loci that constitute our enhancer training set (see Methods). The resulting ROC curve provides a graphical method to distinguish between optimal and suboptimal models in their diagnostic ability. The curve is plotted as the rate of true positives against false positives at given intervals, and the departure of the optimal model from the unity line is taken as a measure of its performance.

To plot the ROC curve, for each of 10 cross-fold validations, we computed Spearman's rank correlations (ρ) between ChIP-seq tag

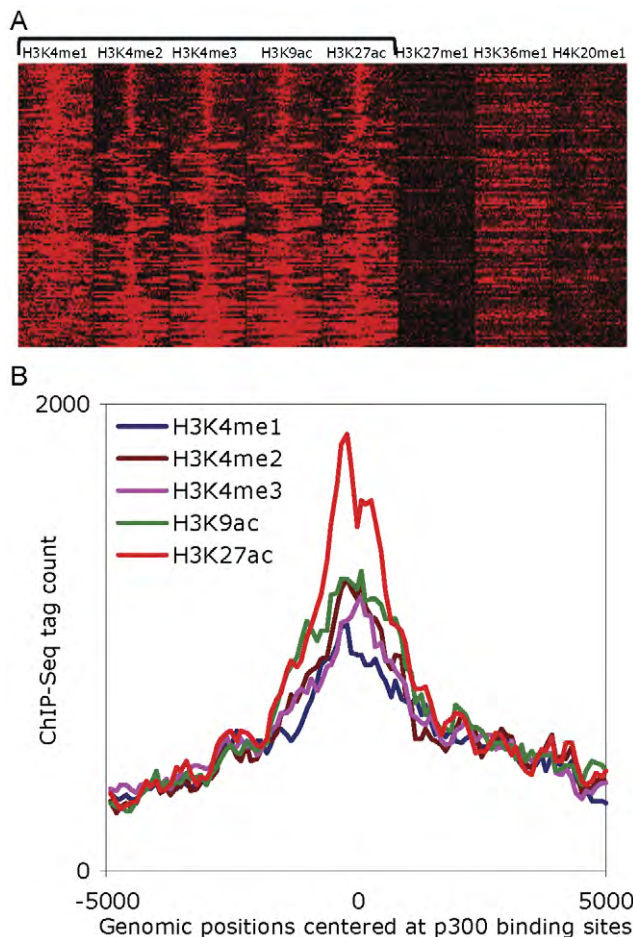


Figure 1. An enhancer training set based on histone modification enrichment. The enhancer training set is derived from five histone modifications in 10 KB windows over 137 p300 binding sites in the K562 cell line. (A) Heat map showing ChIP-Seq tag counts at 137 p300 binding sites for eight histone modifications. The first five of the modifications are significantly enriched and display distinctive patterns at p300 binding sites, whereas the last three modifications do not show any specific pattern over p300 binding sites. (B) Visual representation of the enhancer training set with ChIP-seq tag counts summed over 137 p300 bound genomic loci corresponding to the five enhancer enriched histone modifications binned in 100 bp bins over a 10 kb window.
doi:10.1371/journal.pone.0027513.g001

counts that represent the chromatin profile of the enhancer training sets as an ensemble against tag counts for the p300 bound regions that make up the validation sets (true positives). Then, as a control, we performed similar correlations between the chromatin profile of the training sets and a sets of random genomic loci the same size of the validation sets (false negatives). Across a range of p -values, the fraction of true positives represented by the validation set loci was plotted against the fraction false positives from random genomic loci to yield the ROC curve in Figure 2. The plot demonstrates that our enhancer training set is clearly distinct from the one derived from sampling random genomic loci and thereby possesses the discriminating capability essential for its use in enhancer predictions.

Enhancer prediction

Having established the validity of our enhancer training set, we used it to search for regions that display similar chromatin profiles

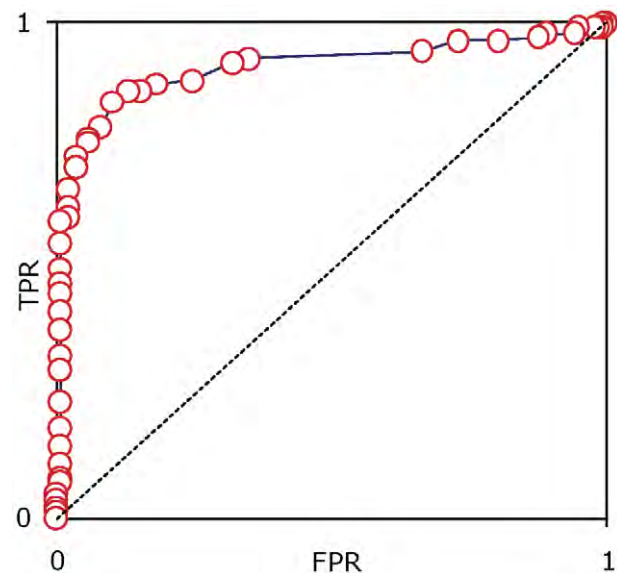


Figure 2. Discriminating ability of the enhancer training set. Receiver operating characteristics (ROC) curve showing the discriminating ability of the enhancer training set. The rate of true positives (TPR) is calculated as the correlations between the chromatin profiles of individual sequences that make up the enhancer training set with the entire enhancer training set considered as an ensemble, and the rate of false positives is calculated as the correlations between the chromatin profiles of randomly sampled genomic loci and the training set (see Methods). Departure of the curve from the unity line is taken as a measure of the discriminating ability of the training set.
doi:10.1371/journal.pone.0027513.g002

in order to identify potential TE-derived enhancers, which may or may not be bound by p300, genome-wide. To that end, we built a test set made up of the DHS sites in the GM12878 and K562 cell lines. Using a 10 kb window and a step size of 100 bp, we computed Spearman's rank correlations between the enhancer training set histone modification profile and test set profiles at each step. For each DHS site, the genomic site that yields the highest correlation value was recorded, and the results were filtered using a correlation cut-off of 0.5 or higher (Spearman's $\rho = 0.5$, $n = 98$, $P = 1E-7$).

Several histone modifications are also known to be enriched at the transcription start sites and promoter regions of human genes [13,15]. Since actively transcribing genes are also associated with DHS, there is a possibility that our enhancer prediction method can potentially misidentify some promoters as enhancers. In order to control for this possibility, we used CAGE (Cap Analysis of Gene Expression) data in each cell line to filter any promoters that may have been identified as enhancers. CAGE tags are obtained by capping the 5' end of messenger RNA and are known to mark the transcriptional start sites of genes [23,24]. We identified loci that are significantly enriched for CAGE tags by using a Poisson distribution parameterized by the background CAGE tag count [16,17]. The potential enhancer predictions that overlapped with CAGE tags were marked as promoters and removed from consideration. After filtering out the promoters in this way, (Delete – CAGE analysis removed) we obtained 11,311 and 8,051 enhancers in the GM12878 and K562 cell lines respectively. A majority of enhancers we identified are unique to each cell line as only 2,114 (10.7%) of these enhancers are shared between both (Figure 3).

Since we limited our search for enhancers to DHS sites, these data reflect the number of enhancers associated with actively

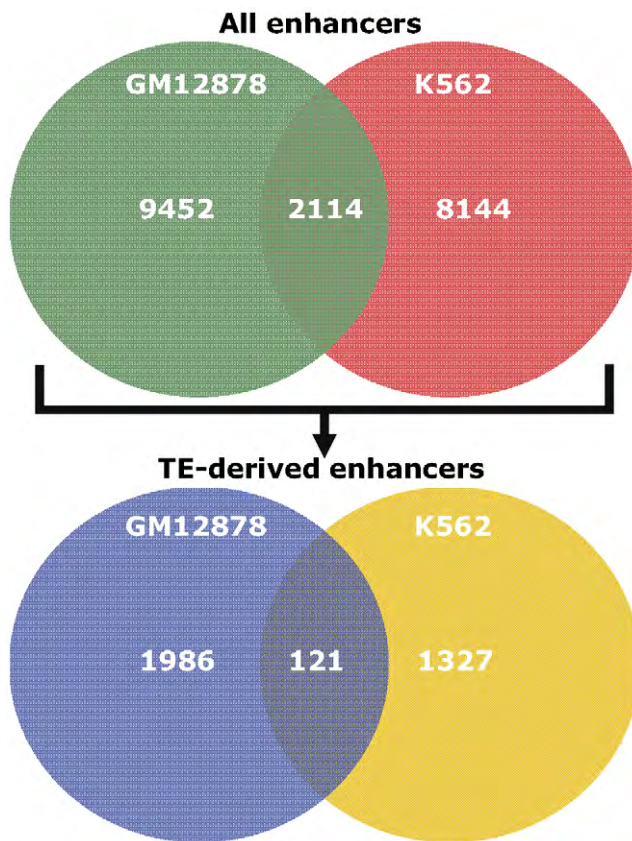


Figure 3. Common and exclusive enhancers between the GM12878 and K562 cell lines. (Top) Venn diagram showing the numbers of enhancers that are shared between the GM12878 and K562 cell lines as well as unique enhancers in the two cell lines, and (bottom) numbers of the enhancers from above that originate in TE sequences are shown.

doi:10.1371/journal.pone.0027513.g003

transcribing genes. Histone modifications in each cell type are dynamic and can change to accommodate the regulatory needs of cell. Thus, the enhancers predicted using the histone modification profiles are also not universally active as reflected by the small percentage of enhancers that are shared between GM12878 and K562 cell lines. As such, these figures provide a snapshot of active enhancers in two human cell lines, each going through a particular stage of differentiation. Accordingly, the divergent genomic loci of these enhancers suggest their role in regulating cell type specific gene expression as discussed in later sections.

TE-derived enhancers

In order to identify functionally active TE-derived enhancers, we intersected the genomic loci of our predicted enhancers in each cell line with TE annotations produced by the RepeatMasker program [25]. We identified 2,107 and 1,448 enhancers derived from TEs in the GM12878 (Table S1) and K562 (Table S2) cell lines respectively with 121 (3.5%) enhancers that are shared between both cell lines (Figure 3 and Table S3). There is a significantly smaller fraction of TE-derived enhancers that are common between the cell lines compared to all predicted enhancers (Hyper-geometric test, $P = 2E-63$), suggesting that TE-derived enhancers are more cell type specific.

To evaluate the contribution of various families of TEs in donating enhancers, we divided TE-derived enhancers into 6 major families, based on the Repbase classification system [26,27],

namely Alu, L1, LTR, DNA, L2, and MIR (Figure 4A). We also normalized the number of enhancers contributed by each TE family by the family's relative genomic abundance (Figure 4B). In both cell lines, Alu and L1 elements are under-represented, whereas LTR, DNA, L2 and MIR are over-represented TE families that contribute enhancers to the human genome (χ^2 test, GM12878: $P = 9E-272$, K562: $P = 4E-217$ –Table S4, Student's t test, GM12878: $t = 5.6$, $P = 1E-3$, K562: $t = 4.6$, $P = 3E-3$). In absolute terms, LTR elements donate the highest number of enhancers (387) in the K562 and the second highest number of enhancers (383) in the GM12878 cell lines. A number of previous studies have demonstrated that LTRs provide transcription start sites and protein coding sequences to the human genome [28,29]. Thus, our analysis extends what is known regarding the extensive regulatory contributions of LTR elements to the human genome. Our data also indicate that MIR elements contribute the largest number of enhancers relative to their genomic abundance. MIRs represent the oldest family of TEs in the human genome, and their over-representation in donating enhancers indicates that older TEs are more likely to provide regulatory and coding sequences for the host genome [30]. Indeed, the relative age of TE families is directly correlated with the number of enhancers it donates (Figure S3, GM12878: $\rho = 0.94$, $P = 3E-19$, K562: $\rho = 0.89$, $P = 2E-17$). The observation that older TE families donate relatively more enhancers than younger ones suggests that older elements may possess a stronger ability to recruit epigenetic marks, making them more likely to be exapted by the host genome. We have also previously shown that older TEs bear more histone modifications than younger ones and therefore demonstrate a higher potential to be exapted by the human genome [31].

TE-derived enhancers and cell type specific gene expression

To evaluate the functional effect of TE-derived enhancers, we investigated their role in the regulation of gene expression. Enhancers can influence the expression of genes that lie as many as tens-of-thousands of bases away from the transcriptional start site of genes. We determined the functional effect of our predicted TE-derived enhancers by relating them to cell type specific gene expression. To do this, we mapped enhancers to genes by finding enhancers in 100 kb windows surrounding transcriptional start sites.

We analyzed GM12878 and K562 gene expression data characterized by exon array experiments as part of the ENCODE project (see Methods) and calculated the average expression of genes that possess different numbers of TE-derived enhancers in their vicinity. For each gene in our dataset, we searched a window of 100 kb surrounding its transcription start site for TE-derived enhancers and binned the average expression of genes with respect to the number of enhancers they possess. The expression of genes without a TE-derived enhancer is significantly lower than that of genes with one or more TE-derived enhancers in the 100 kb region surrounding their transcription start sites (Students' t test, GM12878: $t = 31.2$, $P = 3E-208$; K562: $t = 31.4$, $P = 4E-211$). Furthermore, the expression level of genes is strongly positively correlated with the number of TE-derived enhancers it has in its vicinity (Figure 5) (Spearman's GM12878: $\rho \sim 1$, $p = 3E-3$, K562: $\rho \sim 1$, $p = 3E-3$). These findings suggest that TE-derived enhancers make a contribution to the up-regulation of the expression of nearby genes. As a control, we did the same analysis using non-TE derived enhancer sequences predicted in the same way; the results are qualitatively identical underscoring the potential functional significance of TE-derived enhancers (Figure S4).

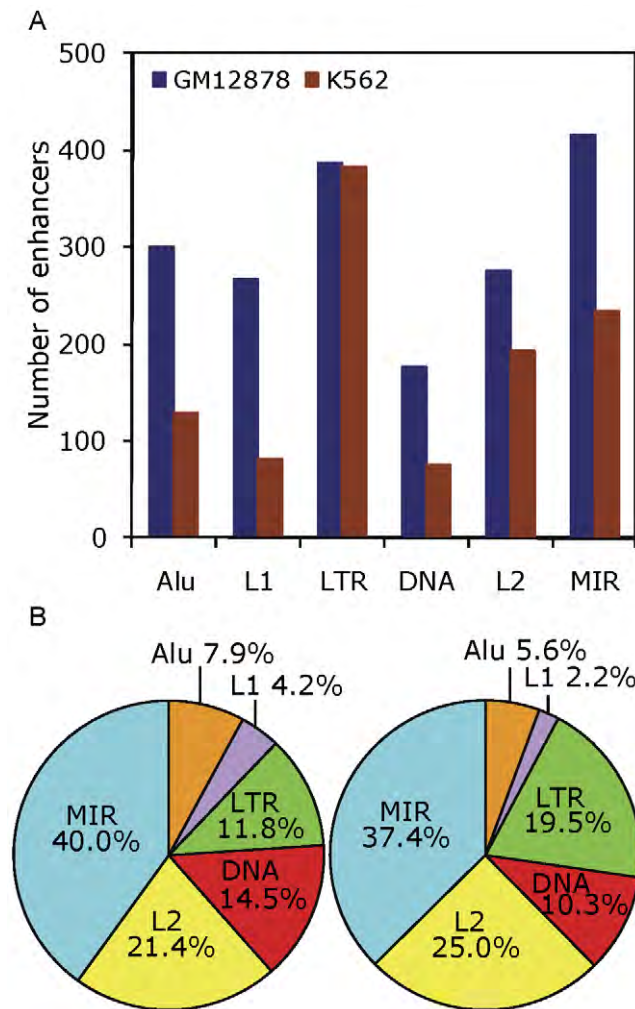


Figure 4. Contribution of various TE families in providing enhancers to the human genome. (A) The number of enhancers provided by six TE families in GM12878 (blue) and K562 (brown) cell lines. (B) Contribution of enhancers by TE families normalized by their genomic abundance (see Methods). doi:10.1371/journal.pone.0027513.g004

TE-derived enhancers and differential expression

Having established the likely functional relevance of TE-derived enhancers in regulating cell type specific gene expression, we evaluated their role in driving differential expression between cell lines. For each gene in our dataset, we computed expression divergence between the GM12878 and K562 cell lines and related it to the difference in the number of gene associated TE-derived enhancers between the two cell lines. We sorted the genes based on their expression divergence and binned them into ten bins according to their increasing expression divergence. We found that expression divergence is directly correlated with the difference in the number of TE-derived enhancers between the GM12878 and K562 cell lines (Figure 6; Spearman's $\rho = 0.89$, $P = 1E-3$). In addition, the trend is not entirely linear; the most extreme bins on either end show the greatest relationship between gene expression divergence and TE-enhancer frequency divergence. This suggests that the strongest influence of TE-derived enhancers is observed for the most differentially expressed genes. Non-TE derived enhancers show a similar, if more linear and stronger, positive correlation between gene expression divergence and enhancer divergence (Figure S5).

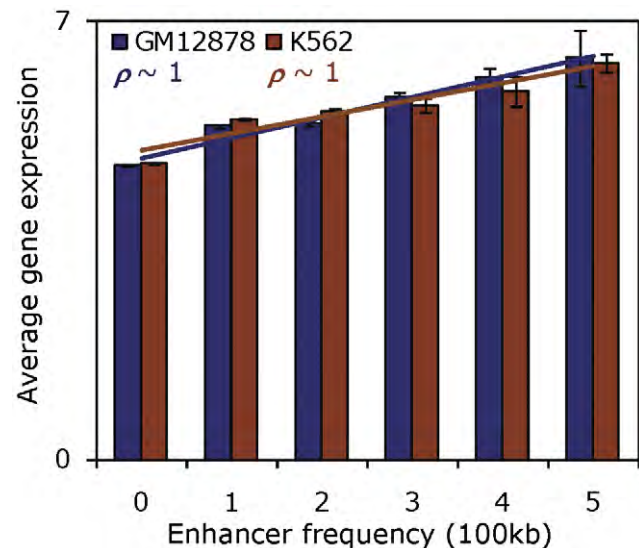


Figure 5. Functional role of TE-derived enhancers in regulating gene expression. Average expression levels (y-axis) of genes that are co-located with different numbers of TE-derived enhancers (x-axis) shown for GM12878 (blue) and K562 (brown) cell lines. doi:10.1371/journal.pone.0027513.g005

In order to further investigate this phenomena, we used ANOVA on 20 and 21 samples of normalized exon array data from the GM12878 and K562 cell lines respectively to determine the maximally differentially expressed genes. We found 4,118 genes that are significantly differentially expressed ($P = 1E-7$) with 1,970 genes that are up-regulated in GM12878 and down-regulated in K562 and 2,148 genes that are down-regulated in GM12878 and up-regulated in K562 cell lines (see Methods). We computed the average number of enhancers in a 100 kb window surrounding differentially expressed genes and found that genes that are up-regulated in one cell line have more enhancers in their vicinity in the same cell line compared to the other cell line (Figure 6). In our dataset of 1,970 genes that are up-regulated in GM12878 and down-regulated in K562, there are an average of 0.43 TE-derived enhancers per gene in GM12878 and 0.17 TE-derived enhancers per gene in K562 cell line (Wilcoxon signed-rank test, $W = 183,472$, $P = 1E-37$). Similarly, the 2,148 genes that are up-regulated in K562 and down-regulated in GM12878 have 0.37 TE-derived enhancers per gene in K562 and 0.30 TE-derived enhancers per gene in GM12878 cell line (Wilcoxon signed-rank test, $W = 143,897$, $P = 4E-4$). These analyses demonstrates that there are more TE-derived enhancers present near genes that are differentially up-regulated in one cell line versus the other, highlighting their contribution to the regulation of differential expression between cell types.

Since we are comparing two cell lines here, it is formally possible that the differences in expression between cell lines are not related to up-regulation of genes associated with cell-type specific TE-derived enhancers in one cell line. Rather, it may be that the corresponding enhancer sequences, which bear distinct chromatin profiles in the alternate cell line, are actually exerting some negative regulatory effect therein. To control for this possibility, we compared the levels of expression for genes associated with cell-type specific TE-derived enhancers against the expression levels of all genes within cell types. We found that the genes associated with cell-type specific TE-derived enhancers are expressed at significantly higher levels than other genes within the same cell type (Student's t-test; GM12878, $t = 18.9$, $P = 2E-73$; K562, $t = 40.0$,

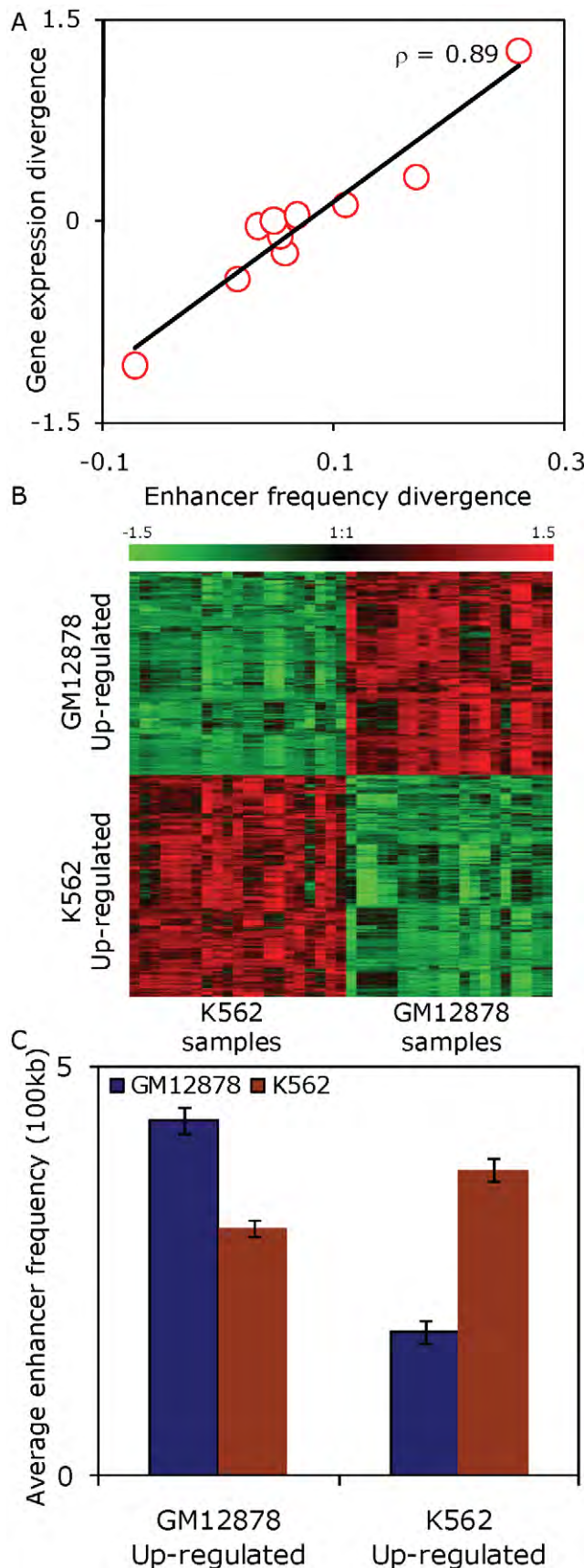


Figure 6. Functional role of TE-derived enhancers in regulating differential gene expression. (A) Gene expression divergence

between GM12878 and K562 (y-axis) is plotted against normalized difference in the numbers of cell type specific TE-derived enhancers (x-axis) co-located with the genes. Expression divergence and enhancer frequency divergence between the GM12878 and K562 cell lines is calculated by subtracting the values of K562 from those of GM12878 cell line. (B) Differentially expressed genes determined by performing ANOVA on the 21 and 20 samples of GM12878 and K562 cell lines respectively (see Methods). (C) The average numbers of co-located TE-derived enhancers found in the GM12878 (blue) and K562 (brown) cell lines are shown for differentially expressed genes that are up-regulated in GM12878 and K562.

doi:10.1371/journal.pone.0027513.g006

$P = 4E-216$). These results are consistent with a positive regulatory role, *i.e.* activation of expression, for the cell-type specific TE-derived enhancers identified here.

Conclusions

Unlike promoters, enhancers can influence the expression of genes that lay tens-of-thousands of bases away from them [32,33]. The distribution of TE derived enhancers around genes ranges from hundreds of bases from the transcription start site to several thousand bases. Enhancers in general have also been shown to provide for the most cell type specific mode of gene regulation [13], and the locations of the TE-derived enhancers we discovered here are even more cell type specific than those of the non-TE-derived enhancers. We used two metrics to investigate the possible functional role of TE-derived enhancers in regulating the expression of genes in a cell type specific manner. Our first analysis revealed that the frequency of TE-derived enhancers in the vicinity of genes is strongly correlated with increasing gene expression. Secondly, genes that are differentially up-regulated in each cell line possess significantly more TE-derived enhancers in the same cell line when compared to the other cell line. These results provide evidence for the functional relevance of TE-derived enhancers in helping to differentially regulate genes between human cell types. Nevertheless, experimental interrogation of individual TE-derived enhancers sequences predicted here will be needed to validate the extent and nature of their regulatory activity. We hope that the list of predicted TE-derived enhancers that results from this work can serve as a guide for further experimental studies on the regulatory contributions of human TEs.

Methods

Enhancer training set and identification of novel enhancers

A dataset of 211 p300 binding sites characterized genome-wide from the K562 cell line was taken to represent functionally active enhancer sequences as previously described [13,14]. We downloaded genome-wide ENCODE histone modification ChIP-Seq data [16,17] for the GM12878 and K562 cell lines from the UCSC Genome Browser for 8 histone modifications characterized in the Bernstein Laboratory at the Broad Institute [34,35,36]: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me1, H3K36me1, H4K20me1. ChIP-Seq tags were mapped to the human genome (UCSC hg18) using the program MAQ with the read-rescue option, which accommodates ambiguous tags that map to multiple genomic regions.

Previously, enhancer locations were predicted using chromatin profiles based on three histone modifications [13]. Here, we have taken a similar approach using additional information afforded by

a total of five enhancer-characteristic modifications as well as DHS sites. DHS sites analyzed were aggregated across 10 kb windows and represent relatively open chromatin; although, they are not entirely devoid of nucleosomes. The p300 binding sites were evaluated for enrichment with five enhancer-characteristic histone modifications: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac. The remaining three histone modifications (H3K27me1, H3K36me1, H4K20me1) were used as negative controls for enhancer regions. A total of 137 p300 binding sites in K562 cells that were found to be significantly enriched for all five enhancer-characteristic histone modifications, but not for enhancer-negative modifications, were used to predict the locations of TE-derived enhancers genome-wide.

For each histone modification, enrichment significance was calculated using a Poisson distribution parameterized by the genomic background ChIP-Seq tag count and the threshold was adjusted using the Bonferroni correction for multiple tests. The enhancer training set was generated using 10 kb windows center-aligned and surrounding 137 p300 binding sites and divided into 100 bins of 100 bp each. Thus, the training set consists of five vectors representing individual histone modifications each made up of 100 bins containing ChIP-Seq tag counts summed over 137 p300 binding sites.

The test set vectors were fashioned in a similar way except in this case individual DHS sites were used instead of p300 binding sites as in the training set. Individual enhancer test set profiles were centered at the start point of the DHS sites and Spearman's rank correlations were computed individually between the five vectors of the training and the test set profiles and the resulting correlations were averaged. We used a sliding window with a step of 100 bp from the start of the DHS sites, computed correlations at every step and took the highest average correlation computed from all the steps within a DHS site. Average Spearman's correlation values of ($\rho = 0.5$, $P = 1E-7$) or higher were taken for further evaluation as potential enhancers.

Cross-validation and receiver operating characteristic (ROC) curve analysis

Ten-fold cross-validation was combined with ROC analysis to evaluate the discriminating power of the enhancer histone modification model. For the 10-fold cross-validation, we partitioned the training set of 137 p300 binding sites into 10 subsamples (with 13 or 14 sequences each), taking 9 of the subsamples as the histone modification training set and a single subsample as the validation set for testing the model. This procedure was iterated 10 times with each of the subsamples used one time as the validation set for testing the model. Then for each cross-validation, Spearman rank correlation coefficients (SCCs) between the histone modification profile of the enhancer training set versus the histone modification profiles of the validation set were computed along with SCCs for the histone modification profile of the training set versus modification profiles of a set of random genomic sequence the same size as the validation set. The resulting ROC curve is based on the relative distributions of the SCC for all ten of the enhancer training versus validation sets (true positives) along with the SCC all ten of the enhancer training versus random genomic sequence sets (false positives). The rates of true positives and false positives were calculated by taking the normalized frequency of correlation values i at regular intervals as described below:

$$TPR_{i=i+0.02}^{-1 < i < 1} = \frac{FoC_{True Positives} > i}{N}$$

$$FPR_{i=i+0.02}^{-1 < i < 1} = \frac{FoC_{False Positives} > i}{N}$$

where interval $i = 0.02$, $N = 137$ (total number of correlation values in each dataset), and FoC = frequency of correlation values in range. The rate of true positives (TRP) was plotted against the rate of false positives (FPR) to yield the ROC curve in Figure 2.

Gene expression analysis

We downloaded Affymetrix exon array signal intensity data from the GEO database under accession number GSE12760. This dataset contains 20 samples of GM12878 and 21 samples of K562 cell lines collected from the different laboratories that are part of the ENCODE project [16,17]. We normalized the dataset using the MAS5 algorithm provided by the Bioconductor package Exonmap [37]. The normalized data was mapped to a genomic locus by averaging the expression values of all probes whose genomic coordinates lay within that the boundaries of that locus for all replicates. We used Refseq genes from the UCSC genome browser to define transcriptional units (TU) [15,38,39]. The TU's, referred to as genes in the text for clarity, encompass the all overlapping co-directional mRNA transcripts at a genomic loci. We defined the boundaries of TUs as the upstream most transcription start site and the downstream most transcription end site.

Differentially expressed genes

Differentially expressed genes were identified using one way ANOVA (Analysis of variance) implemented in the Genesis software package [40]. ANOVA was performed on 20 and 21 samples from GM12878 and K562 cell lines respectively. We used a stringent significance cut-off of $P = 1E-7$ obtained after using Bonferroni correction for multiple tests, to calculate ANOVA.

Sequence annotation datasets

We used five sequence annotation datasets from the March 2006 build (NCBI Build 36.1; UCSC hg18) of the human genome. Three of these datasets were obtained from the ENCODE section of the UCSC Genome Browser [38]. These datasets include histone modifications, DNaseI hypersensitive sites and CAGE data, for both GM12878 and K562 cell lines [16,17]. These data are produced from ChIP-Seq, DNaseI-Seq and CAGE experiments respectively and are available as aligned reads in tagAlign files. Refseq genes and RepeatMasker 3.2.7 data was downloaded from the UCSC Table Browser [39].

Statistical analyses

We used the statistical software R for calculating the Spearman's rank correlation coefficients ρ for all correlation analyses. The statistical significance of Spearman's rank correlation coefficients ρ was determined using the Student's t distribution with $d.f. = n - 2$ with the formula $t = r\sqrt{(n-2)/(1-r^2)}$ [41].

We used a two tailed χ^2 test with $d.f. = 5$ and Student's t test with $d.f. = 1$ to determine the statistical significance of the over- and under-represented TE-families that donate predicted enhancers in each cell line (Figure 4 and Table S4). The genomic abundance of TE families was used to compute the expected number of enhancers derived from each family.

The Wilcoxon signed-rank test was used to establish the statistical significance for the difference in the average number

of enhancers in the vicinity of genes that are differentially expressed between each cell line (Figure 6C).

Supporting Information

Figure S1 Control 1: Relevant versus non-relevant histone modifications. Histone modifications at 137 p300 binding sites in the K562 cell line are shown. The first five modifications were used to build the training set (H3K4me1, H4K4me2, H3K4me3, H3K9ac, H3K27ac), whereas other modification that show no specific pattern of enrichment over the p300 binding sites and were thus excluded from further analysis (H3K9me1, H3K27me3, H3K36me3, H4K20me1). (PPT)

Figure S2 Control 2: Histone modification enrichment patterns at p300 binding sites versus random genomic loci. Epigenetic histone modification levels at 137 p300 binding sites as well as 137 random genomic loci in the K562 cell line are shown. Random genomic loci do not show any discernable pattern of histone modification enrichment compared to p300 binding sites. (PPT)

Figure S3 Over and under-represented TE families in contributing enhancers. Number of TE-derived enhancers observed for different TE families in the GM12878 (blue) and K562 (brown) cell lines normalized by the relative genomic abundances of TE families. (PPT)

Figure S4 Functional role of non TE-derived enhancers in regulating gene expression. Average expression levels (y-axis) of genes that are co-located with different numbers of non TE-derived enhancers (x-axis) shown for GM12878 (blue) and K562 (brown) cell lines. (PPT)

References

- Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284: 601–603.
- Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. *Nature* 284: 604–607.
- Hickey DA (1982) Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* 101: 519–531.
- Kidwell MG, Lisch DR (2000) Transposable elements and host genome evolution. *Trends Ecol Evol* 15: 95–99.
- Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int J Org Evolution* 55: 1–24.
- Hambor JE, Mennone J, Coon ME, Hanke JH, Kavathas P (1993) Identification and characterization of an Alu-containing, T-cell-specific enhancer located in the last intron of the human CD8 alpha gene. *Mol Cell Biol* 13: 7056–7070.
- Yang Z, Boffelli D, Boonmark N, Schwartz K, Lawn R (1998) Apolipoprotein(a) gene enhancer resides within a LINE element. *J Biol Chem* 273: 891–897.
- Santangelo AM, de Souza FS, Franchini LF, Bumashny VF, Low MJ, et al. (2007) Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. *PLoS Genet* 3: 1813–1826.
- Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, et al. (2008) Possible involvement of SINEs in mammalian-specific brain formation. *Proc Natl Acad Sci U S A* 105: 4220–4225.
- Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, et al. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441: 87–90.
- Marino-Ramirez L, Jordan IK (2006) Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol Direct* 1: 20.
- Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* 57: 159–197.
- Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459: 108–112.
- Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39: 311–318.
- Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, et al. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 40: 897–903.
- Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799–816.
- Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, et al. (2010) ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res* 38: D620–625.
- Hatzis P, Talianidis I (2002) Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Mol Cell* 10: 1467–1477.
- Wang Q, Carroll JS, Brown M (2005) Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol Cell* 19: 631–642.
- Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 7: 29–59.
- Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457: 854–858.
- Xi H, Shulha HP, Lin JM, Vales TR, Fu Y, et al. (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet* 3: e136.
- Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, et al. (2006) CAGE: cap analysis of gene expression. *Nat Methods* 3: 211–222.
- Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, et al. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* 100: 15776–15781.
- Smit AFA, Hubley R, Green P RepeatMasker Open-3.0.
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110: 462–467.
- Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7: 474.

Figure S5 Functional role of non TE-derived enhancers in regulating differential gene expression. Gene expression divergence between GM12878 and K562 (y-axis) is plotted against normalized difference in the numbers of cell type specific non TE-derived enhancers (x-axis) co-located with the genes. Expression divergence and enhancer frequency divergence between the GM12878 and K562 cell lines is calculated by subtracting the values of K562 from those of GM12878 cell line. (PPT)

Table S1 1,986 TE-derived enhancers in the GM12878 cell line. (TXT)

Table S2 1,127 TE-derived enhancers in the K562 cell line. (TXT)

Table S3 121 TE-derived enhancers shared between the GM12878 and K562 cell lines. (TXT)

Table S4 χ^2 statistics for over and under represented TE families in contributing enhancers. (PPTX)

Acknowledgments

The authors would like to thank Jianrong Wang for helpful discussions and technical advice.

Author Contributions

Conceived and designed the experiments: AH ET LM-R IKJ. Performed the experiments: AH ET NJB DJ. Analyzed the data: AH ET LM-R NJB DJ IKJ. Contributed reagents/materials/analysis tools: AH ET LM-R NJB DJ IKJ. Wrote the paper: AH IKJ.

28. Cohen CJ, Lock WM, Mager DL (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* 448: 105–114.
29. Feschotte C (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9: 397–405.
30. Silva JC, Shabalina SA, Harris DG, Spouge JL, Kondrashov AS (2003) Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet Res* 82: 1–18.
31. Huda A, Marino-Ramirez L, Jordan IK (2010) Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mob DNA* 1: 2.
32. Machon O, van den Bout CJ, Backman M, Rosok O, Caubit X, et al. (2002) Forebrain-specific promoter/enhancer D6 derived from the mouse *Dach1* gene controls expression in neural stem cells. *Neuroscience* 112: 951–966.
33. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. *Science* 302: 413.
34. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120: 169–181.
35. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 125: 315–326.
36. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448: 553–560.
37. Miller CJ, Okoniewski MJ, Yates T (2007) Description of exonmap: simple analysis and annotation tools for Affymetrix exon arrays.
38. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res* 31: 51–54.
39. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32: D493–496.
40. Sturn A, Quackenbush J, Trajanoski Z (2002) Genesis: cluster analysis of microarray data. *Bioinformatics* 18: 207–208.
41. Sokal RR, Rohlf JF (1981) *Biometry: The Principles and Practice of Statistics in Biological Research*. San Francisco: W. H. Freeman.

Analysis of Biological Features Associated with Meiotic Recombination Hot and Cold Spots in *Saccharomyces cerevisiae*

Loren Hansen^{1,2}, Nak-Kyeong Kim³, Leonardo Mariño-Ramírez^{1,4*}, David Landsman¹

1 Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, **2** Boston University, Bioinformatics Program, Boston, Massachusetts, United States of America, **3** Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia, United States of America, **4** PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia

Abstract

Meiotic recombination is not distributed uniformly throughout the genome. There are regions of high and low recombination rates called hot and cold spots, respectively. The recombination rate parallels the frequency of DNA double-strand breaks (DSBs) that initiate meiotic recombination. The aim is to identify biological features associated with DSB frequency. We constructed vectors representing various chromatin and sequence-based features for 1179 DSB hot spots and 1028 DSB cold spots. Using a feature selection approach, we have identified five features that distinguish hot from cold spots in *Saccharomyces cerevisiae* with high accuracy, namely the histone marks H3K4me3, H3K14ac, H3K36me3, and H3K79me3; and GC content. Previous studies have associated H3K4me3, H3K36me3, and GC content with areas of mitotic recombination. H3K14ac and H3K79me3 are novel predictions and thus represent good candidates for further experimental study. We also show nucleosome occupancy maps produced using next generation sequencing exhibit a bias at DSB hot spots and this bias is strong enough to obscure biologically relevant information. A computational approach using feature selection can productively be used to identify promising biological associations. H3K14ac and H3K79me3 are novel predictions of chromatin marks associated with meiotic DSBs. Next generation sequencing can exhibit a bias that is strong enough to lead to incorrect conclusions. Care must be taken when interpreting high throughput sequencing data where systematic biases have been documented.

Citation: Hansen L, Kim N-K, Mariño-Ramírez L, Landsman D (2011) Analysis of Biological Features Associated with Meiotic Recombination Hot and Cold Spots in *Saccharomyces cerevisiae*. PLoS ONE 6(12): e29711. doi:10.1371/journal.pone.0029711

Editor: I. King Jordan, Georgia Institute of Technology, United States of America

Received: November 28, 2011; **Accepted:** December 1, 2011; **Published:** December 29, 2011

Copyright: © 2011 Hansen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine and National Center for Biotechnology Information. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: marino@ncbi.nlm.nih.gov

Introduction

Meiosis is the biological process by which the genome is divided in half to generate daughter cells that can participate in sexual reproduction. In eukaryotes, this process is accompanied by meiotic recombination, which involves pairing of homologous chromosomes and exchanging of genetic material. Meiosis serves to increase genetic diversity in progeny (for review see [1] and [2]). Recombination does not occur with a uniform frequency across the genome. Instead, there are regions with high and low recombination rates called hot and cold spots, respectively. Recombination is initiated by double-strand breaks (DSBs) which are catalyzed by Spo11 [3]. In this biological event, broken DNA ends are processed to produce single-strand ends that can invade the homologous chromosome [4].

Mapping DSB hot spots [5,6,7] and factors correlated with hot/cold spot formation is an active area of research. Several biological features have been found to correlate with higher levels of Spo11-catalyzed DSBs. Genome-wide mapping and analysis of Spo11-catalyzed DSB sites in the yeast *Saccharomyces cerevisiae* showed that regions with a high break frequency had a high G+C content [7]. A recent study using this same dataset revealed that several types of microsatellites were associated with recombination hot spots [8].

Additionally, studies using machine learning-based techniques and sequence-based features have differentiated DSB hot and cold spots somewhat successfully [9,10], suggesting that differences in sequence composition between these regions exist.

In addition to sequence-based factors, chromatin structure is associated with regions of high and low recombination. Many hot spots exhibit an open chromatin structure constitutively in both meiotic and mitotic cells [11,12]. Some of these hot spots also show an increase in micrococcal nuclease (MNase) sensitivity in meiotic cells shortly before DSB formation [13], indicating active chromatin remodeling to a more open configuration upon the onset of meiosis. Some posttranslational histone marks are also associated with increased DSB frequency, with H3K4me4 and bulk histone acetylation (in *Schizosaccharomyces pombe*) showing a positive correlation [14,15] and H3K36 methylation exhibiting a negative correlation. Here we used a multivariate feature selection approach to determine the sequence and chromatin features that best distinguish hot and cold spots in *S. cerevisiae*. The histone modifications and nucleosome occupancy data used in our analysis were derived from vegetatively growing mitotic cells, which is a different cell state than meiotic cells. Genome-wide epigenetic studies using both mitotic and meiotic states were used to increase the amount of useable data; there is good reason to believe that

epigenetic marks found at hot or cold spots in mitotic cells will also be present at those same sites in meiotic cells (see Discussion).

Feature selection is a dimensionality reduction technique designed to identify the subset of features that is most informative in producing robust predictive models. Feature selection has been used successfully in microarray gene expression studies [16,17] and biomarker identification [18,19]. When attempting to build a classifier based on vectors of features, many features are irrelevant. For example, a common task in microarray studies is to identify which genes are relevant in distinguishing between two or more experimental conditions. In this case, the expression level of thousands of genes (i.e., features) is measured, but only a small subset is relevant in discriminating between the experimental conditions. Many pattern recognition techniques were not designed to deal with circumstances in which the number of relevant features is outnumbered by irrelevant ones [20]. In these instances, feature selection can be used to reduce over fitting, improve predictive performance by identifying a subset of relevant features, and provide insight into the underlying biological processes that generated the data. Machine learning-based approaches have already been applied to the problem of discriminating between hot and cold spots [9,10]. However, these studies analyzed low resolution data and feature selection was not performed. Here we report the results of applying feature selection to identify factors associated with recombination hot and cold spots. A feature vector as used in this study is a string of numerical features; each feature in the string represents a measurement of a biological quantity.

Methods

Definition of hot and cold regions

Buhler *et al.* [5] mapped the frequency of meiotic DSBs in *S. cerevisiae* with high resolution tiling arrays. Using this data, we obtained 1179 and 1028 regions identified as hot and cold spots, respectively, for a total of 2207 regions. Each region was 600 base pairs (bp) in length. Buhler *et al.* produced a set of peaks representing hot spots with 5-fold and 2-fold enrichment over background. In our analysis, hot spots were defined by centering a 600-bp window at the midpoint of peaks that were enriched 5-fold over background. Cold spots were obtained by finding at least three adjacent probes with a log₂ hybridization ratio of less than 0.75, and then centering a 600-bp window at the midpoint of the centermost probe. For each region, we produced a vector of length 350 to represent features such as the chromatin-associated factors “Nucleosome occupancy”, “H3K14ac”, “H3K36me3”, “H3K4me1”, “H3K4me2”, “H3K4me3”, “H3K79me3”, and “H3K9ac”.

Pan *et al.* [21] identified hot spots by mapping the binding of Spo11 using high throughput sequencing. We centered 600-bp windows at the middle of hot spots as defined by Pan *et al.* Cold spots were defined by a set of non-overlapping 600 bp windows with no reads aligned that did not overlap to any extent simple repeats as downloaded from the UCSC genome browser.

Generation of chromatin structure-based features

Pokholok *et al.* used tiling arrays to map histone modifications in *S. cerevisiae*. We obtained this data from the public database ArrayExpress Archive (<http://www.ebi.ac.uk/arrayexpress/>) and normalized using MA2C normalization [22]. There are a number of publically available datasets containing additional chromatin marks mapped genome wide that potentially could have been included in this study. Unfortunately they are low resolution; one microarray element per ORF or intergenic region or they do not control for differences in nucleosome occupancy. For each region,

we obtained the degree of enrichment by averaging the normalized hybridization values of the probes within that region. For example, the feature “H3K14ac” represents the average degree of acetylation of lysine 14 in histone H3 for the given region. A similar approach was used for each histone modification. To calculate the degree of nucleosome occupancy we used a dataset produced by Kaplan *et al.* [23]. For most positions in the genome, Kaplan and co-authors calculated a nucleosome occupancy score. The average nucleosome occupancy was normalized to zero. A value greater than zero represents nucleosome enrichment relative to the genome-wide average, while a value less than zero signifies nucleosome depletion. For each hot or cold region, nucleosome occupancy was calculated by averaging the nucleosome occupancy scores for that region.

Generation of sequence-based features

In this study, 342 out of 350 features were sequence-based in which each sequence feature represented the normalized frequency of the region for one of the 1–4 possible k-mers. For example, feature 9 for region \times would be the number of times the 2-mer “AT” was found in the region divided by the number of k-mers of size 2 found in the region. Hence the feature represents the enrichment of AT relative to all 2-mers found in the region. Similarly, feature 300 for region \times would be the number of times “AAGT” was found in the region divided by the number of k-mers of size 4 found in the region. We also included two sequence features “AT content” and “GC content”, reflecting the overall AT and GC content for that region, respectively. It would seem the sequence features could further be reduced by removing the reverse complement of the given k-mer (CG is the same as GC). Whether or not the reverse complement is redundant is based on whether or not strand specific processes are acting at Hot spots. There are examples of strand specific trans-acting factors operating at hot spots [24]. Hence reverse complements were retained in the final set of features.

Feature selection

Feature selection can be described as finding the subset of features from the set of all possible combinations of features that can best distinguish classes of interest. Because the search space of all possible combinations of features grows exponentially with the number of features, it is rarely feasible to perform an exhaustive search. Instead, various heuristic search methods can be used to identify meaningful feature subsets that can be used to build classifiers with high accuracy. Here we used a genetic algorithm (GA)- based approach [25] similar to those published previously [26,27,28]. We used the R package Galgo [29] to implement the algorithm.

The dataset of 2207 features was divided randomly into two groups, a training dataset containing 1471 regions and a testing dataset containing 736 regions. Each dataset contained roughly equal numbers of hot and cold regions. The training dataset was further divided into three pairs of sub-training and validation datasets. Each pair of the sub-training datasets contained 981 regions, while those of the validation dataset contained 490 regions. The GA was then applied to these datasets in search of a subset of features with optimal accuracy based on the average accuracy across all sets of sub-training and validation data. More specifically, the GA searched for a feature subset that optimized a score defined as $A_{\text{total}} = (A_1 + A_2 + A_3)/3$, in which A_i is defined as the accuracy of the given subset of features using a random forest classifier built utilizing the sub-training dataset i and tested on the validation dataset i and $i = \{1, 2, 3\}$. In general, accuracy was defined as the total number of regions classified correctly divided

by the total number of regions in the validation dataset. The search space of 350 features was prohibitively large, and running the GA twice on the same training and validation datasets would most likely have yielded two different solutions representing local optima. Thus, a sampling of the fitness landscape was used in which the GA was run 10,000 times on different random divisions of the training dataset into the sub-training and sub-validation datasets. The final solution was obtained by combining the results of these independent runs. Features were ranked according to their frequency of occurrence within the subset of optimal features selected by the GA. Features that were present across many runs were presumed to be more important than those that were selected less often (Figure 1). For example, if feature one was present in 9,000 of the 10,000 optimal subsets returned by the GA, while feature two was present in only 5,000, then feature one would be considered more important and thus ranked higher than feature two. The final subset of features was obtained using a forward selection approach. Features were added individually based on ranking until no significant improvement in accuracy was observed. The corresponding accuracy was calculated using the testing dataset.

Alignment methodology

Alignments were performed using BLASTN with default parameters [30]. When allowing multimapping of reads we followed the procedure as defined in [31]; briefly any alignment yielding an identity less than 90% was discarded and, for alignments between 90% and 95%, only the maximum score was retained. All alignments with greater than 95% identity were kept. Identity was defined as alignment length divided by read length.

MNase control subtraction methodology

Normalizing for differences in sequencing coverage was accomplished by dividing read counts at each base pair by the total number of unique mappable reads for each dataset, similar to the procedure used in [32]. The following formula was used to subtract out the normalized counts of the MNase control. Given two sequencing datasets D_1 and a control D_2 with normalized counts of read coverage at each base pair represented by $c_1 = \{c_{1,1}, c_{1,2}, \dots, c_{1,m}\}$ and $c_2 = \{c_{2,1}, c_{2,2}, \dots, c_{2,m}\}$, the subtracted read density was defined at each base pair as

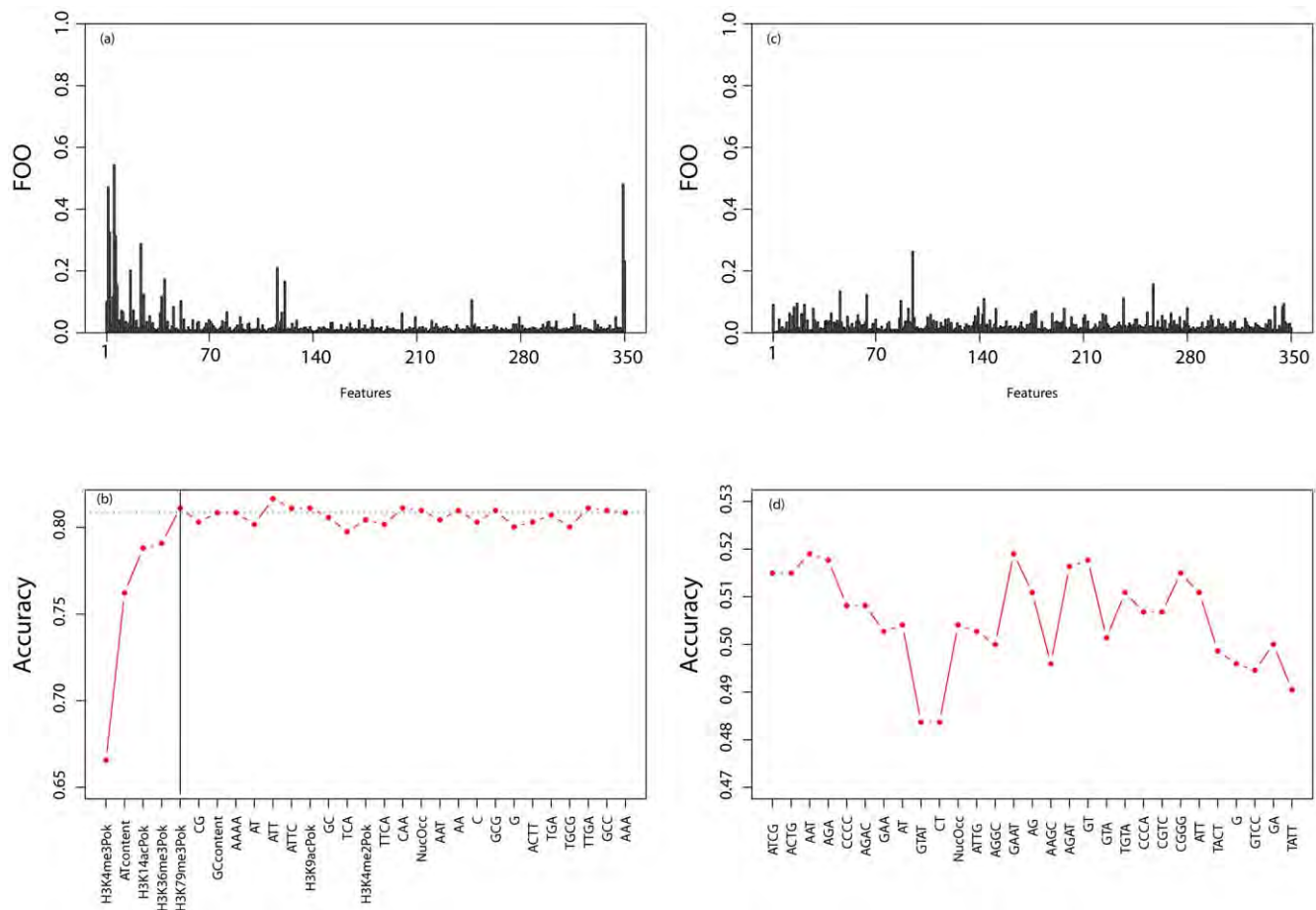


Figure 1. Overview of the feature selection procedure. The initial set of 2207 regions was divided into a training set of 1471 regions and a testing dataset containing 736 regions. The training dataset was further divided into sub-training and validation datasets. (a) The (Genetic Algorithm) GA was run 10,000 times on different sub-training and validation datasets, producing a subset of optimal features for each run (see Methods). We divided the number of times each feature occurred in an optimal feature subset by the total number of times the GA was run (i.e., 10,000) to calculate the frequency of observation (FOO). Features that occurred most often in many different optimal subsets across different splits of the training dataset were ranked higher than features that were selected less often. (b) To obtain the final subset, features were added individually based on their FOO score from highest to lowest. Then, the corresponding accuracy using the testing dataset was calculated. Features were added until no substantial improvement in accuracy was observed, indicated in the figure panel (b) by the solid black line. Panels (c) and (d) are identical to (a) and (b) except random regions were used (i.e., 1179 and 1028 regions randomly selected and labeled as “hot” and “cold”, respectively). doi:10.1371/journal.pone.0029711.g001

$$c(i) = \log \left(\frac{c_{1,i} + \alpha}{c_{2,i} + \alpha} \right)$$

where α is a constant set to 2 to avoid division by zero errors and to dampen noise.

Results

A dataset consisting of 2207 regions (1179 hot spots, 1028 cold spots) was first randomly divided such that two-thirds were analyzed by feature selection (see Materials and Methods) and one-third was set aside as a testing dataset. The testing dataset was used to test how accurately the features identified can distinguish between hot and cold spots. Setting aside a testing dataset ensures a fair test with the features being tested on data not used to obtain the features. Hot spots as used in this manuscript refer to regions of increased meiotic DSBs and cold spots to regions of decreased meiotic DSBs. Features were first ranked in order of importance based on the training dataset. The final subset of features was obtained using a forward selection approach. Features were added individually based on ranking until no significant improvement in accuracy was observed. The corresponding accuracy was calculated using the testing dataset. Thus, accuracy using only highly ranked features was estimated based on data not used to rank the features. A subset of five features (i.e., H3K4me3, H3K14ac, H3K36me3, H3K79me3, and GC content) was identified (Figure 1) with a classification accuracy of 80.4%, sensitivity of 80.5%, and specificity of 80.3%. Many of the identified features were found to be associated with recombination, according to published literature.

Chromatin Structure

All of the histone modifications used in this study were mapped in vegetatively growing mitotic cells. While the DSB frequency dataset used to map meiotic hot and cold spots was obtained from meiotic cells we address this issue in more detail in the discussion section. The feature selected as having the highest predictive importance was the degree of H3K4me3 methylation. Published literature strongly associates this mark with recombination hot spots. In *S. cerevisiae*, the methyltransferase Set1 is responsible for H3K4 methylation. Set1 mutants exhibit dramatically reduced DSB frequency at well-characterized hot spots [33]. Additionally, H2B ubiquitination promotes Set1 activity [34], thereby increasing H3K4 methylation. Preventing this mark leads to decreased DSB frequency [35]. Importantly, Borde *et al.* [14] demonstrated that deleting Set1 reduced or eliminated DSBs at 84% of the hottest sites in *S. cerevisiae*. In addition, recent work has associated PRDM9, a sequence-specific DNA binding methyltransferase, with hot spot activity in mammalian meiosis [36,37,38]. Our results are consistent with these studies, indicating that H3K4me3 associates positively with areas of high recombination (Figure 2).

H3K14ac is a histone mark associated with active transcription. Like H3K4me3, H3K14ac is localized primarily to the 5' end and promoter region of open reading frames and is correlated with the rate of transcription [39,40,41]. Research has linked histone acetylation with meiotic DSB frequencies. For instance, Sir2 deacetylates histones H3 and H4 [42]. Mutants deficient in Sir2 exhibit widespread changes in meiotic DSB frequencies with 12% of yeast genes showing altered DSB frequency [43]. Moreover, the histone deacetylase Rpd3 represses meiotic recombination at the well-studied hot spot *HIS4* in *S. cerevisiae* [44]. Finally, deletion of the histone acetyltransferase *GCN5*, which preferentially acetylates H3 histones, leads to decreased recombination at the *ade6-M26* hot

spot in *S. pombe* [15]. Our analysis indicates that H3K14ac is associated with DSB hot regions, with high levels of this mark corresponding to hot spots and low levels to cold spots (Figures 2 and 3).

H3K36me3 is a post-translational modification catalyzed by the methyltransferase Set2, and is found primarily in the coding region of genes being actively transcribed [39,40]. By recruiting the repressor Rpd3, H3K36me3 suppresses spurious transcription initiation [45]. H3K36me3 may also play a role in differentiating exons from introns [46]. Our results indicate that the presence of H3K36me3 may play a largely inhibitory role in DSB frequency as this mark is enriched in cold spots relative to hot spots (Figures 2 and 3). In addition, studies have shown that Set2 the methyltransferase responsible for H3K36me3 represses meiotic recombination at the *HIS4* hot spot in yeast [44].

Like H3K36me3, H3K79me3 is found primarily within coding regions. Unlike H3K36me3, however, the degree of H3K79me3 presence is not strongly associated with transcription [40]. The exact function of this mark is unknown, although some evidence suggests that H3K79me3 may play a role in histone H3 exchange [47]. Our results indicate H3K79me3 may play a minor repressive role in DSB frequency since cold spots appear to be enriched for H3K79me3 (Figures 2 and 3). Most of the histone modification features show a strong partitioning with hot spots being either enriched or depleted for the chromatin mark and vice versa for cold spots. H3K79me3 is an exception cold spots are enriched for this mark but hot spots are not depleted instead showing about the genome average of H3K79me3 (Figure 2 panel a). This trend could be explained by H3K79me3 having a lesser effect on DSB frequency or by an indirect effect.

Computational analysis is rarely capable of demonstrating a causal relationship. Feature selection can identify which biological features out of a large number of candidate features are associated with regions of high/low meiotic DSBs. The method cannot identify the reason behind the association. Once an association is discovered it is important to identify potential confounding variables and test whether they may be solely responsible for the correlation of biological features. Such an analysis cannot prove a causal relationship but it is helpful in elucidating uninteresting correlations.

An important confounding variable that arises when working with recombination hot spots is their tendency to localize to promoter regions while cold spots localize to coding regions. Many of the histone marks we studied also have a tendency to localize either to the 5' end of genes or to coding regions. Therefore, it is possible that the results of our analysis reflect this co-localization effect. To explore this, we compared promoter regions of genes with a hot spot within 500 bp upstream of the transcription start site (TSS) (N = 218) to those genes whose TSS is at least 3000 bp away from a hot spot (N = 2491) (Figure 3 panels a and d). Divergent promoters were removed from this analysis. Gene coordinates were obtained from the UCSC genome browser.

Both H3K14ac and H3K4me3 exhibit a "peak" of modification in promoters of genes that contain hot spots. This "peak" is absent in promoters that lack hot spots. H3K14ac and H3K4me3 are positively correlated with transcription. It is possible that the enrichment of H3K14ac and H3K4me3 observed upstream of genes close to hot spots is due to increased transcriptional rates. To test this we obtained gene expression data [48] and compared transcription rates. The set of genes with a hot spot upstream of the TSS, on average do have a higher transcriptional rate compared with genes whose TSS is at least 3000 bp away from a hot spot (2.2 mRNA/h compared to 1.7 mRNA/h, p-val-

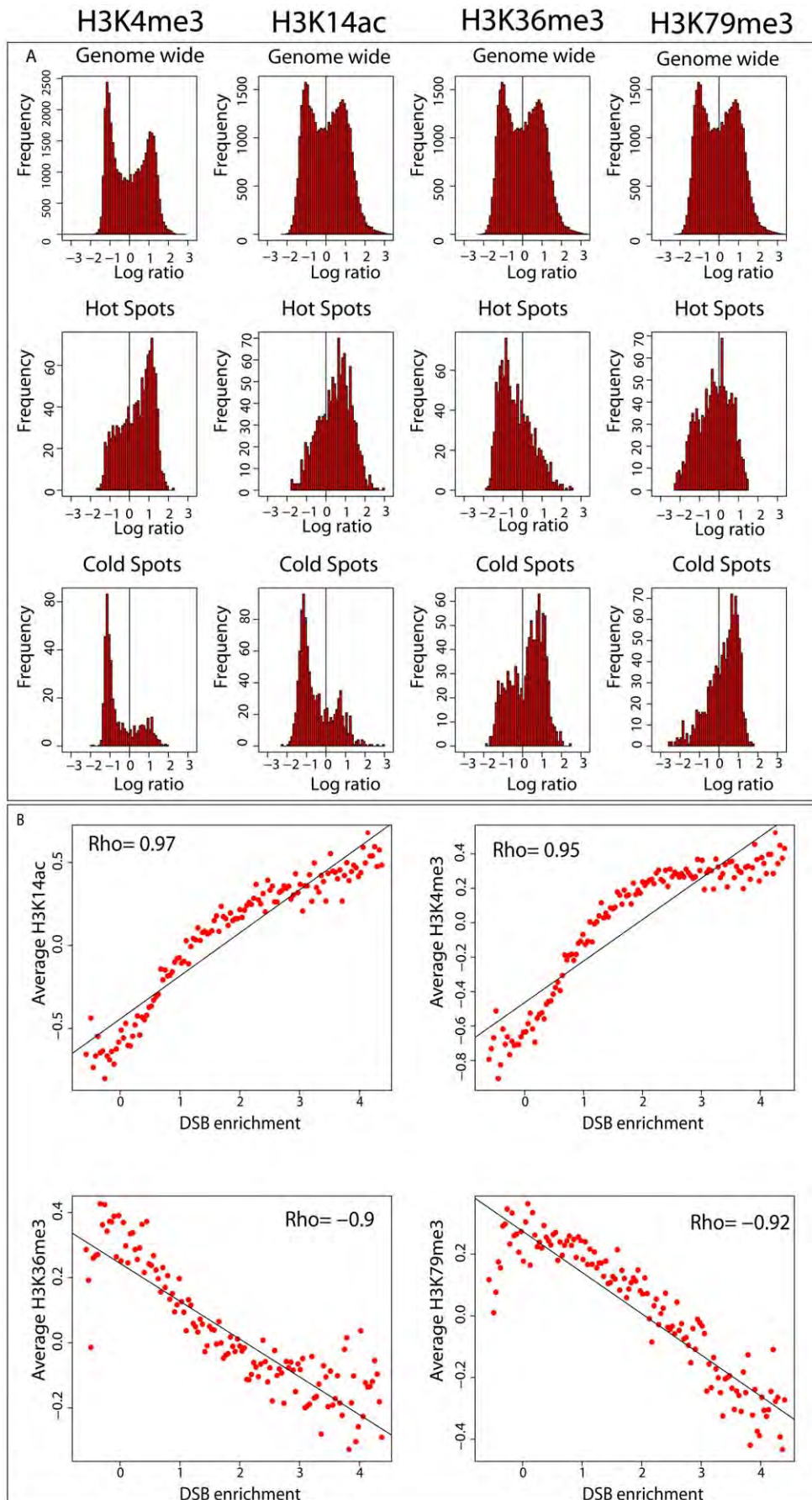


Figure 2. Selected histone marks are correlated with meiotic DSB frequency. (A) Presence of histone marks at hot or cold spots. The first row displays histograms of the log ratios for all probes on the microarray. The higher the log ratio, the more enriched is the given mark. The second row is the enrichment of the histone marks at hot spots. Log ratios were binned in 600-bp windows centered at hot spots and the averages for each bin plotted. The third row is the enrichment of the histone marks at cold spots. Log ratios were binned in 600-bp windows centered at cold spots and the averages for each bin plotted. (B) Histone mark enrichment is correlated with DSB frequency. Probes on both microarrays measuring DSB enrichment and histone modification were paired based on whether they mapped to the same genomic location. Pairs of probes were then grouped in 100 bins according to their DSB enrichment (x-axis). The corresponding log ratios measuring histone modification for the given mark were then averaged for the probes in each bin (y-axis). Bins representing extreme DSB enrichment values had a very low number of probes ~1–10 hence the histone modification averages for these bins was highly variable. Therefore any bin containing less than 50 probes was discarded.
doi:10.1371/journal.pone.0029711.g002

$ue = 0.003$, Wilcoxon rank sum test), an association that has previously been reported [7].

To test whether this difference in transcription could explain the extra enrichment of H3K14ac and H3K4me3 upstream of the TSS we plotted these marks for genes with an upstream hot spot whose transcriptional rate was less than 1 mRNA/h (Figure 3 panels b and e) ($N = 43$). The peaks of upstream enrichment are retained even for inactive genes. This analysis indicates that H3K14ac and H3K4me3 enrichment in areas of high recombination is likely not due solely to the tendency of hot spots to localize to promoter regions or to differences in transcriptional activity. Similarly, we compared coding regions that entirely contain a cold spot to those that do not overlap to any extent with cold spots (Figure 3, panels c and f). Genes that contain cold spots show an increased enrichment for both H3K36me3 and H3K79me3. Both H3K36me3 and H3K79me3 within gene bodies are positively correlated with transcriptional activity [40], H3K36me3 is strongly correlated and H3K79me3 is weakly correlated. Perhaps the increased enrichment of H3K36me3 and H3K79me3 in genes containing cold spots compared to genes without cold spots is due to the fact that cold spots are preferentially located in active genes. We compared transcriptional rates for genes with ($N = 498$) and without cold spots ($N = 4516$). Genes with cold spots have lower transcription rates than genes without cold spots (median transcriptional rate 1.3 mRNA/h compared to 2.3 mRNA/h, $p\text{-value} < 1e-16$ Wilcoxon rank sum test). Even though genes containing cold spots have on average lower transcriptional rates than genes without cold spots they exhibit a higher degree of H3K36me3 and H3K79me3 methylation (Figure 3 panels c and f).

Holstege *et al.* measured gene expression in mitotic cells. The purpose behind the preceding analysis is to check whether the observed patterns of histone modifications at hot or cold spots are due to differences in gene activity and not to the presence or absence of a hot or cold spot. Given that the histone modifications were measured in mitotic cells, the appropriate dataset for the above analysis is gene expression also measured in mitotic cells. While this manuscript was in preparation, a high resolution map of DSB hot spots was published [21]. This map was produced by sequencing and mapping oligos bound by Spo11 where the hot spots were mapped at much higher resolution than the Buhler *et al.* dataset. We obtained the set of hot spots mapped by Pan *et al.* in order to check if the association of meiotic DSB frequency with the histone marks H3K14ac, H3K4me3, H3K36me3 and H3K79me3 observed using the Buhler *et al.* dataset were also observed using an independently produced higher resolution hot spot map. The Pan *et al.* hot spots, like the Buhler *et al.* hot spots, strongly localized to promoter regions [21]. Hence, a positive correlation with H3K14ac and H3K4me3 and a negative correlation with H3K36me3 and H3K79me3 would be expected.

We duplicated the analysis described in Figure 3 using the Pan *et al.* hot spots, and found similar results to what was seen using the Buhler *et al.* hot spots. Additionally, we show that the H3K14ac and H3K4me3 peaks observed upstream of genes with a hot spot

are in general proportional to the strength of the hot spot (Figure S1). The comparison of gene expression rates between hot spot associated genes and non-hot spot associated genes and cold spot associated genes with non-cold spot associated genes was performed using gene expression obtained in vegetatively growing mitotic cells. To check if the same patterns are observed with meiotic cells we repeated the above comparisons with gene expression measured at different time points after cells were placed in sporulation media (Figure S2) gene expression data was taken from [49]. The expression dataset used measured gene expression for four yeast strains SK1, non-sporulating SK1 control, W303 and a non-sporulating W303 control. For the non-sporulating controls which do not enter meiosis the above described patterns held true for all time points. That is hot spot associated genes are transcriptionally more active than non-hot spot associated genes and cold spot associated genes are transcriptionally less active than non-cold spot associated genes.

Interestingly, this pattern did not hold true in the case of hot spot genes compared to non-hot spot genes in meiotic cells. Upon the entrance to meiosis the difference in gene expression between hot and non-hot genes gradually falls to zero (Figure S2 panel's b and d). This could be explained by the observation that hot spot associated genes have a tendency to be repressed in meiosis [7]. Cold spot associated genes are transcriptionally less active than non-cold spot genes in both mitotic and meiotic cells (Figure S2 panel's e, f, g and h).

As discussed above there is ample evidence from multiple studies that H3K4me3 is involved in hot spot selection. Given that histone marks are in general correlated with one another [39], is it possible the association of H3K14ac, H3K79me3, and H3K36me3 with DSB frequency is simply a consequence of these marks being correlated with H3K4me3? In the case of H3K36me3 there is previous research linking this mark with hot spot activity at a well-studied hot spot in yeast [44]. As discussed above multiple studies have linked histone acetylation with hot spot activity.

H3K4me3 in general is correlated with other histone marks but it is particularly strongly correlated with H3K14ac ($r = 0.85$, $p\text{-value} < 2.2 \times 10^{-16}$) compared to its correlation with H3K4me2 which is the next strongest correlation ($r = 0.62$, $p\text{-value} < 2.2 \times 10^{-16}$). Even when comparing a large number of histone marks H3K4me3 is inordinately strongly correlated with H3K14ac [39]. Taken together with the previous work linking histone acetylation with recombination, the usually strong correlation of H3K4me3 with H3K14ac combined with our results suggests these marks may act together at meiotic DSB hot spots. While there is a statistically significant correlation between H3K4me3 and H3K79me3 ($r = 0.09$, $p\text{-value} < 2.2 \times 10^{-16}$) this correlation is too small and in the wrong direction to explain the association of H3K79me3 with meiotic DSB frequency.

AT/CG Content

One of the features selected by the feature selection algorithm was a sequence based feature AT content. AT content and GC content measure the same quantity and both were included in the

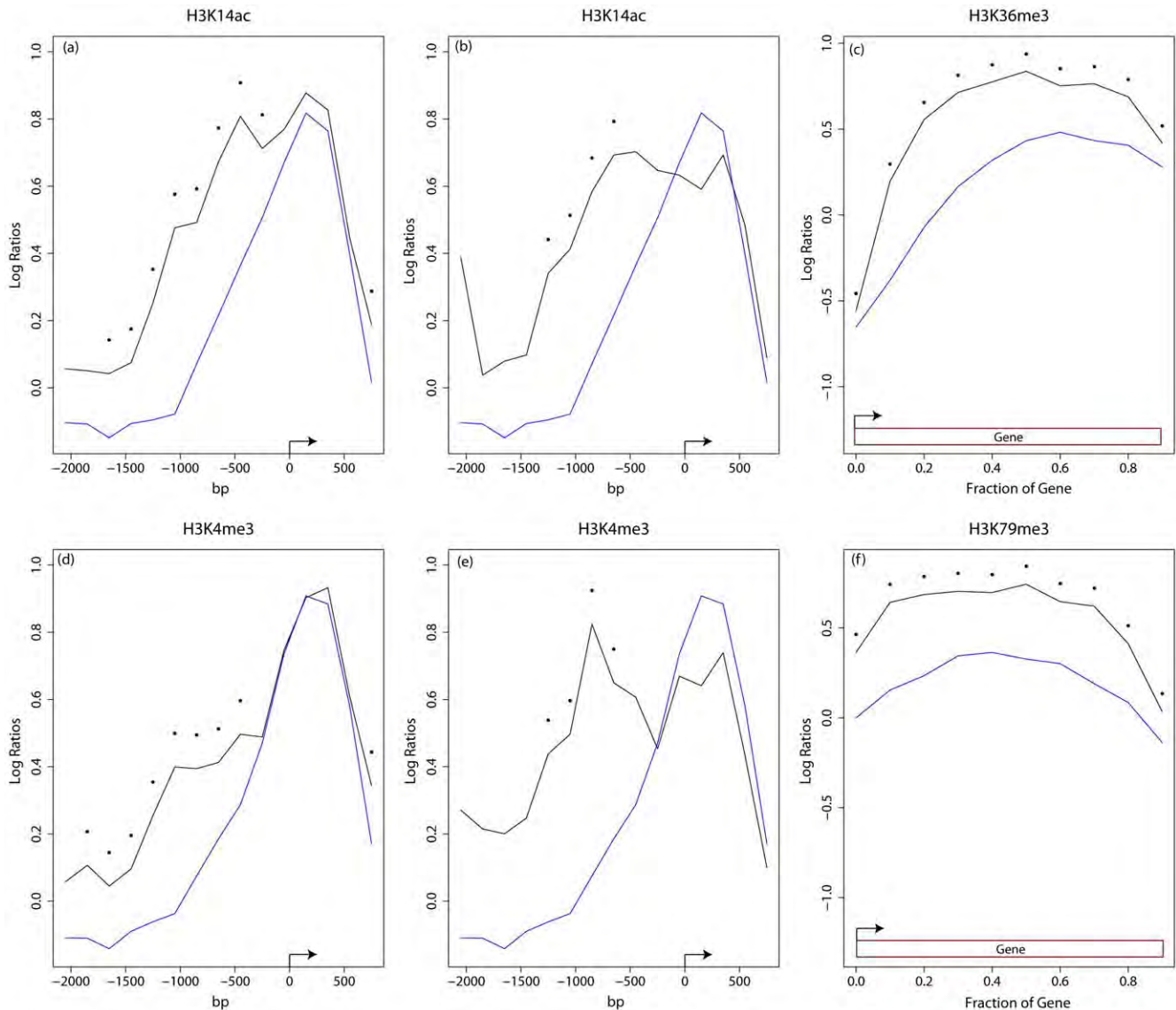


Figure 3. Plots of average modification level around transcription start sites (TSS). The x-axis represents position relative to the TSS set at zero. Positive numbers represent positions downstream of the TSS, while negative numbers are upstream. The y-axis indicates the average histone modification enrichment log ratios. Black dots represent points statistically significantly different (p -value < 0.01 wilcoxon rank sum test) than the corresponding point in the other curve. For panels (a, b, d and e) the blue line represents TSS at least 3000 bps away from the center of a hot spot, log ratios were binned in 200-bp windows and the average for each bin plotted. The black line represents genes with the center of a hot spot located within 500 bp upstream of the TSS, log ratios were binned in 200-bp windows and the average for each bin plotted. For panels (c and f) the black line represents the average histone modification in genes which entirely contain a cold spot (for definition of cold spot see Methods). The blue line represents the average histone modification in genes which do not overlap to any extent a cold spot. Plots were produced by binning histone modification log ratios in bins proportional to gene size (each bin was 1/10 the size of the given gene) the average for each bin is plotted.
doi:10.1371/journal.pone.0029711.g003

input feature set as a “sanity check” or control. If our computational method is working correctly, then these features should rank similarly. Indeed, this is what was observed AT content ranks 2nd out of 350 features GC content ranks 7th (Figure 1). Our analysis is in agreement with published results [7] indicating that GC content in hot spots is higher than the overall average in *S. cerevisiae*. More specifically, the mean GC content within a 600-bp window centered on hot spots was 39.6%, while the GC content of the entire genome was 38.1%. Not surprisingly, the mean AT content in cold spots (63.8%) is greater than that across the entire genome (61.9%).

To further explore the relationship between GC content and recombination cold spots we examined the set of cold spots found entirely within coding sequences. Coding sequences in yeast have a GC content of 39.6%, which is GC rich relative to the genome as a whole. The mean GC content of cold spots found entirely within coding sequences was 37.0% compared to the genome average of 38.1% and compared to 36.0% percent GC content calculated for the entire set of cold regions. Cold spots found within otherwise GC-rich regions (i.e., coding sequences) still showed reduced GC content contrary to the overall trend of coding regions as a whole. Studies have shown that hot spots are

generally absent from protein coding sequences despite their high GC content [50,51]. Our results suggest that cold spots may be associated with regions of low GC and high AT content within coding sequences.

Nucleosome Occupancy

Of the four biological features included in our analysis with previous evidence from the literature associating them with meiotic DSBs (H3K4me3, H4K36me3, GC content and nucleosome occupancy) three were selected by our method (H3K4me3, H3K36me3, GC content). Our method did not identify nucleosome occupancy as an important feature distinguishing hot from cold spots. This is surprising since multiple studies [11,12,52] have suggested that recombination hot spots are typically found in regions of increased sensitivity to nucleases, presumably reflecting a local open chromatin structure. The dataset we used to test nucleosome occupancy was produced by Kaplan *et al.* [23] and based on high throughput sequencing technology.

One possible explanation for our results is that chromatin remodeling may be occurring after cells have entered meiosis. Kaplan *et al.* measured nucleosome occupancy using data derived from vegetatively growing mitotic cells. There are examples of hot spots showing a closed chromatin structure during mitosis but an open one in meiosis [53]. However, a recent study that measured nucleosome occupancy using formaldehyde-assisted isolation of regulatory elements (FAIRE) showed that meiotic DSB hot spots

genome-wide overlapped with nucleosome-free regions in mitotic cells greater than would be expected by random chance [54] which greatly weakens the above hypothesis. To investigate this further, we obtained a set of nine different nucleosome occupancy maps from three microarray-based and six high throughput sequencing-based studies and examined nucleosome occupancy around hot spots in each dataset. All six sequencing-based datasets fragmented DNA using nuclease digestion. Two of the microarray-based nucleosome positioning maps used sonication. One of them, (Figure 4 (c)) similar to the sequencing-based datasets used micrococcal nuclease digestion [55]. The Lee *et al.* dataset also mapped nucleosome positions at a high resolution ~ 4 bp similar to the 1 bp resolution of the sequencing-based studies. Our analysis yielded a discrepancy in the results comparing microarray- and sequencing-based nucleosome occupancy maps. The microarray-based results all show a well-defined valley representing nucleosome depletion centered at hot spots. Based on these results and previously referenced studies, we conclude that the microarray results best approximate what occurs *in vivo*. On average, nucleosomes are depleted at hot spots for mitotically dividing cells. Contrary to these results, the sequencing-based datasets yielded a small peak of nucleosome occupancy at hot spots (Figure 4). Some datasets exhibited a variable amount of bias (compare peak to baseline differences Figure 4 panels d and e to Figure 4 panel f).

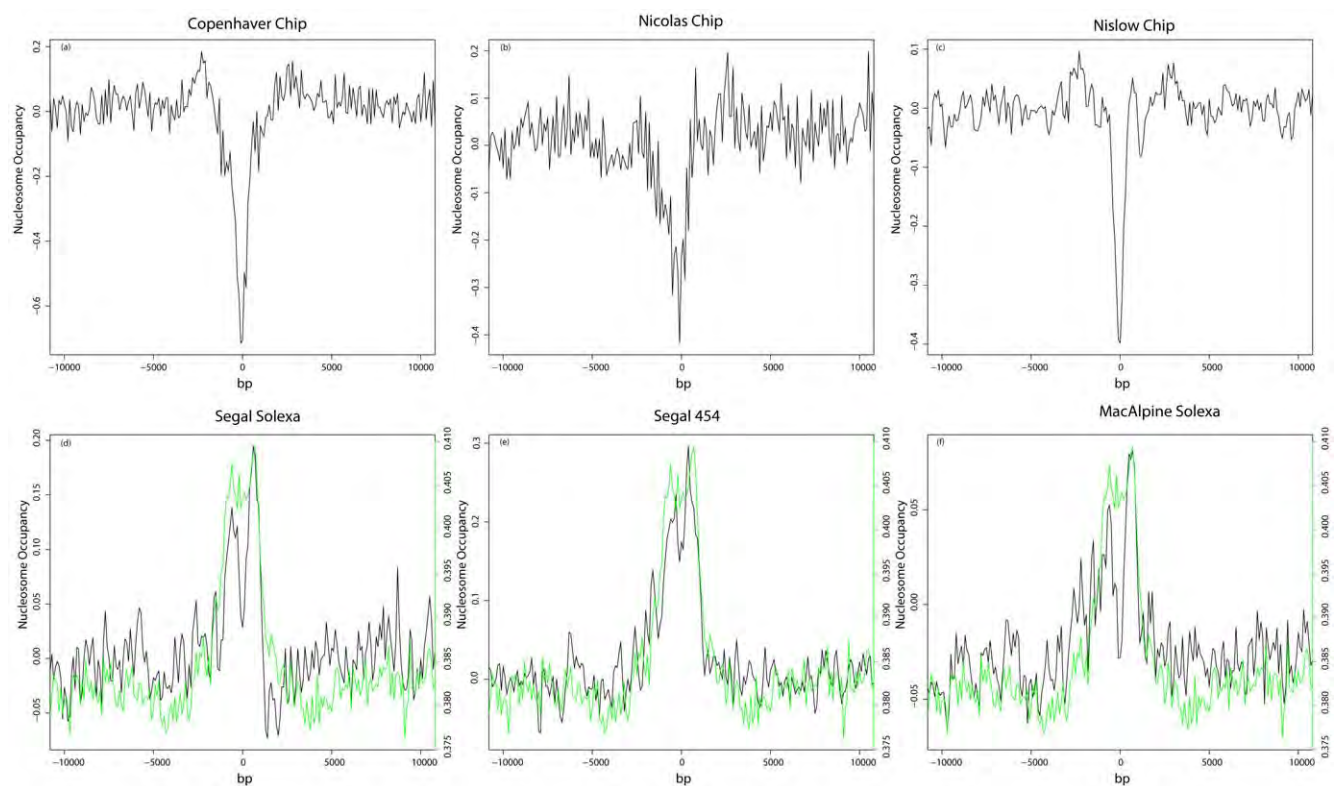


Figure 4. Nucleosome occupancy at hot spots. Multiple nucleosome occupancy maps produced using three different technologies (i.e., FAIRE, Chip-Chip, Chip-Seq) were obtained. Hot spots were aligned Z-score standardized nucleosome occupancy as is shown in 100 bp bins (y-axis). The center of the aligned hot spots is zero on the x-axis. (a–c) Nucleosome occupancy maps based on microarray technology. The sign was reversed in panel a to be consistent with how nucleosome depletion is represented in the other microarray-based techniques. (d–f) Nucleosome occupancy maps based on high throughput sequencing. The green line plots the mean GC content around hot spots as calculated by averaging the GC content in 100-bp bins. The y-axis scale on the right is for the GC content plot. The first word in each plot title is the last author on the paper in which the given dataset was described. (references for datasets: a [54], b [14], c [55], d [23], e [76], and f [77]). Nucleosome occupancy scores were used as calculated by the authors.

doi:10.1371/journal.pone.0029711.g004

We obtained and plotted read density at and around hot spots using two publicly available control datasets (Figure 5). Control dataset “a” was produced by micrococcal nuclease (MNase) digestion of purified DNA followed by size selection for nucleosome-sized fragments and subsequent sequencing using the Solexa platform [56]. Control dataset “b” was the product of sonicated purified DNA followed by size selection for nucleosome-sized fragments and sequenced using the Solexa platform [57]. Both control datasets showed a peak of read density at hot spots very similar to the peak of nucleosome occupancy observed in the six sequencing-based nucleosome occupancy maps implying nucleosome occupancy at hot spots, as measured by high throughput sequencing, is likely dominated by experimental artifacts. Because the read density peak was observed in both controls, this bias was most likely not introduced by a MNase sequence preference.

The nucleosome occupancy maps produced using high throughput sequencing show a split peak with a small valley of occupancy centered at hot spots. The low point of this valley is still higher than or equal to the baseline nucleosome occupancy (Figure 4 panels d, e and f). This split peak is likely due to the competing influences of depleted nucleosome density at hot spots with the peak of control read density also centered at hot spots. Thus the trend observed with the sequencing datasets at hot spots is the result of experimental bias as seen in the control datasets combined with nucleosome depletion as seen in the microarray results.

A recent study [21] mapped hot spots and nucleosome occupancy in yeast at high resolution using high throughput sequencing, showing nucleosome depletion at hot spots. Using this

dataset we plotted read density at and around hot spots for the MNase and the sonication controls. Similar to the results seen for the Buhler *et al.* hot spots, there is a spurious peak of read density at the Pan *et al.* hot spots (Figure S4). This is likely due to GC content bias, Pan *et al.* hot spots correlate with a higher GC content similar to the Buhler *et al.* hot spots [21]. However, when we plotted nucleosome occupancy at the Pan *et al.* hot spots using the same six sequencing based nucleosome occupancy maps we plotted at the Buhler *et al.* hot spots we observed a valley of nucleosome occupancy centered at hot spots contrary to the peak seen with the Buhler *et al.* hot spots (compare Figure S3 with Figure S5). There is wide variability in the level of bias within the sequencing based nucleosome occupancy datasets examined. This can be seen comparing the distance of the peak height to the baseline in Figure 4 panels d, e and f and Figure S3. The effects of this variability in bias can also be seen when plotting nucleosome occupancy at the Pan *et al.* hot spots (Figure S5). Those datasets with the strongest bias exhibit a strong split peak with depletion centered in the middle of a peak (Figure S5 panel a). Those datasets with a weaker bias show a much smaller split peak (Figure S5 panels c and f).

The Pan *et al.* hot spots are mapped with much higher resolution than the Buhler *et al.* hotspots. A higher fraction of the mapped Pan *et al.* hot spots will be located close to or at the real hot spot, which is likely to be nucleosome depleted; therefore the Pan *et al.* hot spots will have a higher signal to noise ratio than the Buhler *et al.* hot spots. The lower signal to noise ratio of the Buhler *et al.* hot spots is sufficient using microarray based nucleosome occupancy maps, such that the correct biological conclusion can be obtained (Figure 4 panels a, b and c). Using biased nucleosome occupancy

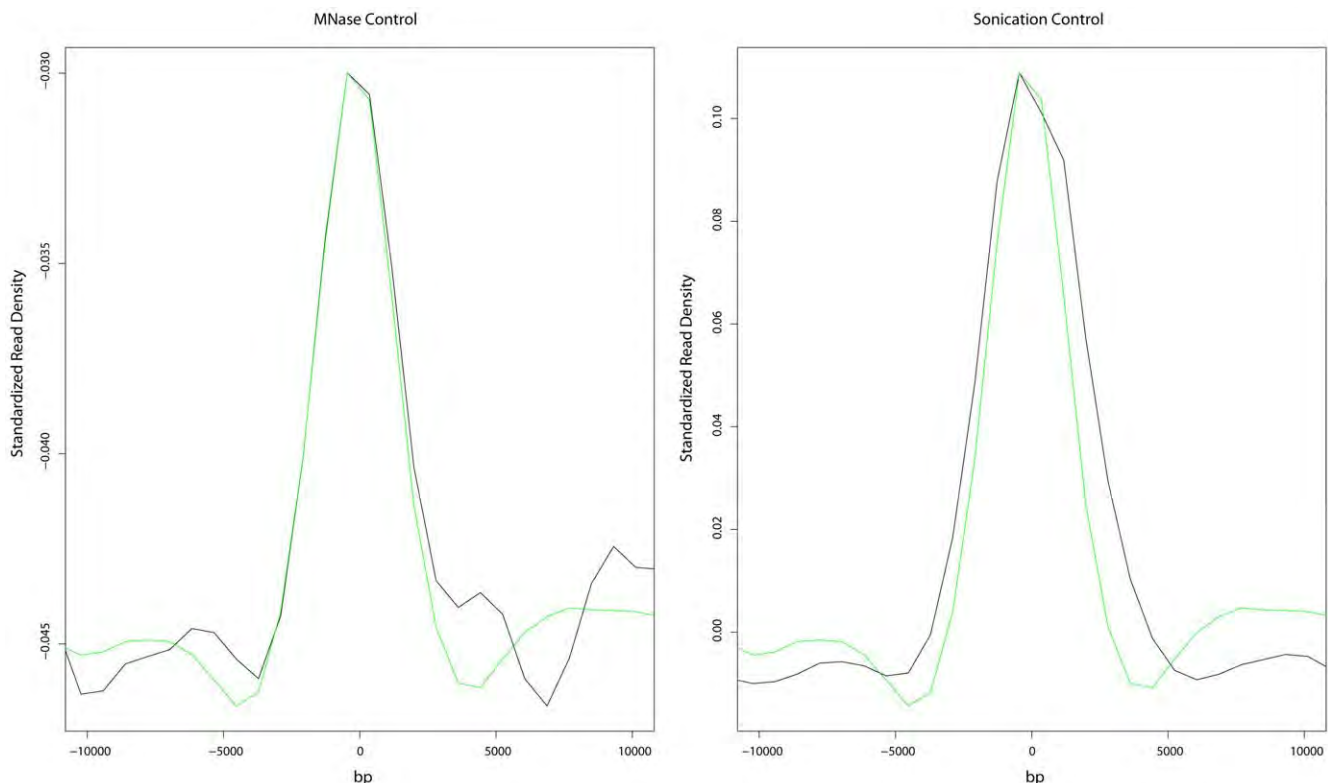


Figure 5. Read density for sequencing controls at hot spots. (a) Purified DNA digested with micrococcal nuclease (MNase) and sequenced using the Solexa platform. (b) Purified DNA following sonication and sequencing using the Solexa platform. The black line indicates the z-score standardized mapped read density, while the green line depicts GC content as calculated in Figure 4. Data was smoothed using loess smoothing. doi:10.1371/journal.pone.0029711.g005

maps the lower signal to noise ratio of the Buhler *et al.* hot spots is not sufficient and an incorrect biological conclusion is drawn (Figure 4 panels d, e and f). These same nucleosome occupancy maps, when used with hot spots mapped with much higher resolution and a corresponding greater signal to noise ratio like the Pan *et al.* hot spots can qualitatively produce the correct biological picture (Figure S5).

To further examine this issue using a single sequencing based nucleosome occupancy map, we plotted nucleosome occupancy at three different hot spot datasets: Buhler *et al.* [5], Borde *et al.* [14] and Pan *et al.* [21]. Depending on which hot spot maps were used, nucleosomes were either depleted at hot spots or nucleosome occupancy at hot spots was more difficult to distinguish from baseline (Figure S6 panels d, e and f). Also plotted is a single nucleosome occupancy as mapped by ChIP-chip [55] for the three different sets of hot spots. Contrary to the sequencing based nucleosome occupancy maps, the ChIP-chip based map showed clear nucleosome depletion regardless of which hot spot datasets were used (Figure S6 panels a, b and c). Using high-resolution hot spot datasets coupled with sequencing based nucleosome occupancy maps supports an accurate qualitative interpretation. However, it is quantitatively difficult to determine nucleosome occupancy due to the bias imposed by the sequencing technologies.

It is tempting to conclude that the bias observed at hot spots is due to a GC content bias in next generation sequencing. Our results, in agreement with others [7] demonstrate that hot spots have a tendency to be GC-rich. Several studies have reported evidence of significant GC content bias in next generation sequencing [58,59,60,61]. In support of this hypothesis, plots of nucleosome occupancy near the center of hot spots closely mirror those of GC content (Figure 4, panels d, e and f, and Figure S3).

To further explore this question, read libraries for all six sequencing-based nucleosome occupancy maps plus two control datasets were aligned against the yeast genome, and the GC content of reads that aligned with at least 95% identity (alignment length divided by read length) was calculated. This set was further divided according to whether the reads mapped to intergenic or coding regions (Table 1). An obvious GC bias was discovered in mappable reads (Table 1, column 4). Studies have shown intergenic regions are nucleosome poor compared to coding regions [55,62]. Since nucleosomes are concentrated to some extent in GC rich coding regions and coding regions are GC-rich a genome-wide examination of sequence bound by nucleosomes would be expected to find a high GC content relative to the genome average. However, it is unlikely that this effect can

completely explain the GC bias shown by the six sequencing-based datasets. The GC content in coding regions of the yeast genome is 39.6% whereas that shown by reads mapped to coding regions is ~42.0%. At 41.3%, the GC content of reads mapped to intergenic regions is much higher than the GC content of intergenic regions (34.8%).

Comparison of the GC bias between the two control datasets was particularly interesting. The MNase control showed a strong GC bias in mappable reads of 47.6%, which was nearly 10.0% higher than the overall yeast GC content. The sonication control displayed a much lower GC content bias (39.2%) for mappable reads. All of the sequencing-based nucleosome occupancy maps were produced using MNase digestion. Given the clear GC bias calculated for the MNase control, it is possible that much of the GC bias shown by these maps is a product of MNase cleavage bias. Furthermore, our analysis indicates that the bias seen at hot spots occurs regardless of sequencing platform. Nucleosome occupancy maps produced using both Solexa and 454 sequencing exhibited a bias at recombination hot spots. Given the differing nature of these sequencing platforms, the bias may be introduced during sample preparation and not by the sequencing technologies themselves.

Not surprisingly, read mapping methodology can also influence downstream analysis. Five of the six sequencing-based datasets and all of the control datasets used only unique aligned reads. However, Mavrich *et al.* [63] used a more lenient mapping approach whereby any alignment yielding an identity less than 90% was discarded and, for alignments between 90% and 95%, only the maximum score was retained [31]. All alignments with greater than 95% identity were kept. The key difference is that their method retained reads that mapped with high confidence to multiple areas along the genome. Using this mapping strategy, a broad shallow valley of read density was observed at hot spots (Figure 6, panel a). When only unique aligned reads from the same dataset were used, a peak of read density similar to that seen with other sequencing-based datasets was seen (Figure 6, panel b). When the control datasets were examined using the Mavrich *et al.* mapping approach, a similar shallow depletion of read density was observed for the sonication control (Figure S7, panel a). The MNase control showed a similar shallow depletion, with the exception of a small peak of read density centered at hot spots. This peak closely mirrors the increase in GC content also centered on hot spots and is likely due to the increased GC bias seen in the MNase control (Table 1). Hence, depending on the mapping approach, opposing biases can be introduced.

Table 1. Average GC content for reads mapped to the yeast genome.

Dataset	Intergenic GC content	Coding GC content	Total GC content
Yeast Genome	34.84%	39.62%	38.15%
Segal 454	42.35%	42.60%	42.40%
Segal Solexa	41.69%	42.20%	41.97%
Pugh 454	41.49%	42.60%	41.81%
Rando Solexa	39.66%	42.26%	41.52%
Friedman Solexa	41.20%	42.99%	42.46%
MacAlpine Solexa	41.71%	42.96%	42.54%
MNase Control	47.84%	47.37%	47.51%
Sonicated Control	38.58%	39.66%	39.19%

doi:10.1371/journal.pone.0029711.t001

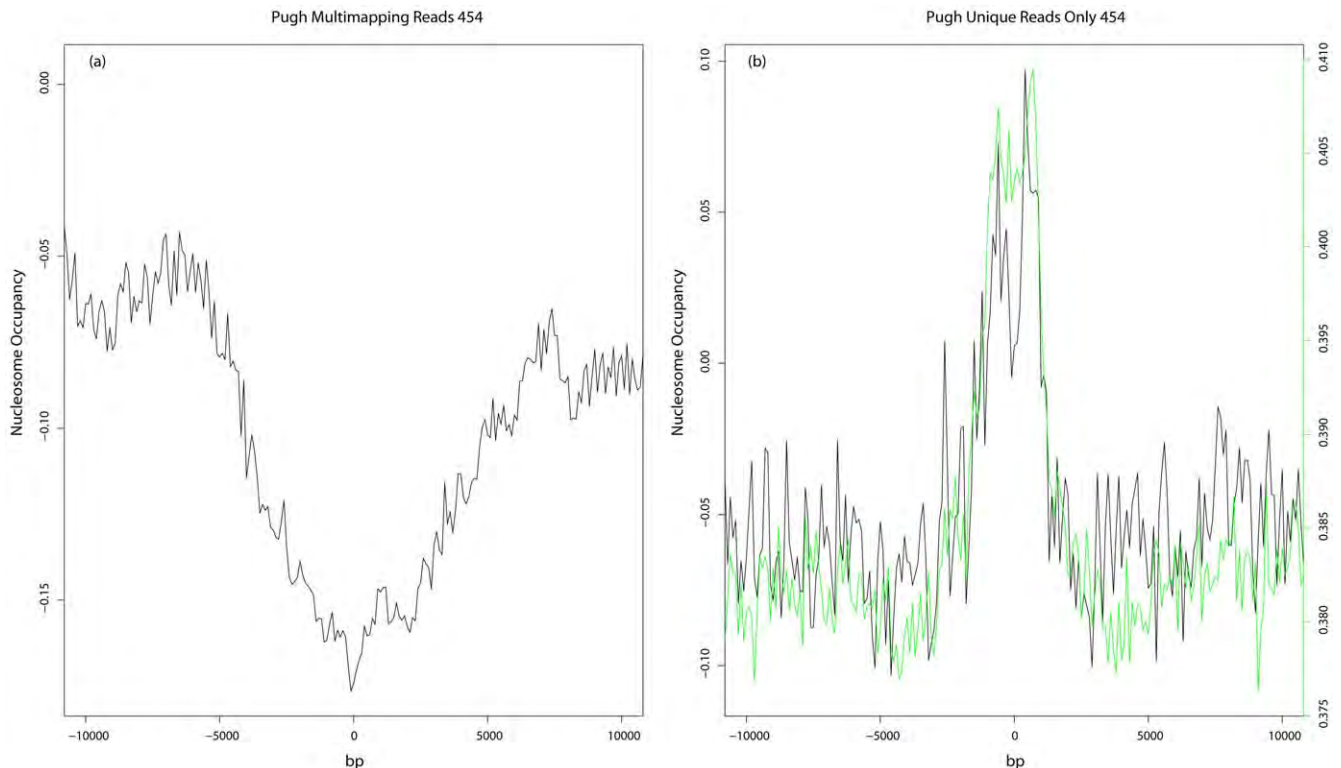


Figure 6. Effect of including multimapping reads. (a) Plot of nucleosome occupancy at hot spots using data produced by the Mavrich *et al.* mapping approach. (b) Plot of nucleosome occupancy of the same dataset at hot spots using uniquely aligned reads only. Green line represents GC content as calculated in Fig. 4.

doi:10.1371/journal.pone.0029711.g006

Using uniquely aligned reads will bias mapped read density towards unique sequence; including multimapping reads will bias read density towards repetitive sequences. The broad shallow depletion in read density observed at hot spots when allowing multi mapping of reads may reflect the fact that hot spots have a tendency to be located in unique sequences.

Next, we plotted the read density for the six sequencing-based nucleosome occupancy maps following subtraction of the MNase control (see Materials and Methods). When the MNase control was subtracted from the nucleosome occupancy maps, the read density at hot spots is qualitatively in agreement with the microarray-based results, displaying a valley of nucleosome occupancy at hot spots (see Figure S8).

Discussion

It is difficult using *in silico* analysis alone to demonstrate the existence of a causal relationship between two biological features. What it can do is to identify promising relationships to explore further *in vivo*. Here we have shown that feature selection using machine learning techniques can usefully be applied to a complex biological process. While this manuscript was in preparation a high resolution map of DSB hot spots was published [21]. Sequencing and mapping oligos bound by Spo11 produced this map. Spo11 hot spots compared with hot spots identified by ssDNA hybridization studies such as Buhler *et al.* show a strong degree of concordance with Spo11 hot spots accounting for nearly all hot spots mapped by ssDNA techniques [21].

Resolution of Hot Spots

The set of DSB hot and cold spots used in this study were derived by mapping single stranded DNA produced by nucleolytic

processing of DSBs [5]. These ssDNA fragments may be quite large, 1 to 2 kb. Hence the locations of hot spots as reported by Buhler *et al.* are mapped with some imprecision. This will certainly affect any study that attempts to use this data to elucidate genomic features associated with DSB hot/cold spots.

It is not necessary in this computational analysis for the sites defined as hot spots to exactly overlap the “true” hot spots. It is only necessary that an appropriately sized window centered at the sites defined as hot spots overlap to some degree with the genomic features that are associated with true hot spots. A recent paper studying the association of H3K4me3 with meiotic DSB found enrichment of this mark in a broad region ~1–2 kb around DSBs [14]. This indicates that regions of high DSB frequency mapped by Buhler *et al.* are likely sufficiently precise to identify at least some chromatin features associated with regions of high meiotic DSBs. Our results strengthen this conclusion of the five features we associated with meiotic DSBs. Three of them H3K4me3, H3K36me3 and GC content have previously been associated with meiotic DSBs. Additionally we obtained a set of recently produced hot spots mapped at high resolution [21] and tested whether the same patterns identified using the low resolution Buhler *et al.* dataset are present using higher resolution data. The same patterns were present using either dataset compare (Figure 3 with Figure S1).

Mitotic Histone Marks

All of the histone marks associated with recombination in this study were obtained in vegetatively growing mitotic cells. The DSB set we used was mapped in meiotic cells. How can we be sure the histone marks do not change dramatically between these two cell states? There are two major reasons suggesting patterns in

histone modifications found at hot spots in mitotic cells may hold true for meiotic cells. First, it has previously been shown that a number of chromatin features present at hot spots in meiotic cells are also present at hot spots in mitotic cells [14,54]. For example, H3K4me3 does not change dramatically in mitotic compared to meiotic cells [14]. The set of hot spots mapped in meiotic cells by Buhler *et al.* have been shown to be on average nucleosome depleted in mitotic cells [54] indicating that at least two chromatin features associated with recombination hot spots in meiotic cells are also present to some degree at those same sites in mitotic cells. Additionally, a recent study examined the changes in chromatin states from mitotic to meiotic cells for a number of nucleosome associated biological features including H3K9ac, H3K4, H3K36 and H3K79 tri-methylation. The conclusion reached was that histone modification states were remarkably stable changing little between mitotic and meiotic cells [64]. These authors also examined the distribution of H3K36me3, H3K4me3 and H3K79me3 at hot and cold spots in meiotic cells. Their results mirror our own obtained in mitotic cells. In addition, Zhang *et al.* showed that in general the distribution of these marks change little between mitotic to meiotic cell states suggesting that the chromatin features associated with hot or cold spots are present in both mitotic and meiotic cells.

Second, we show that, in general, histone modifications peak heights for H3K14ac and H3K4me3 found in promoter regions of genes with hot spots are proportional to the strength of the corresponding hot spots and not dependent on transcriptional rates. The fact that this pattern is present in mitotic cells is strongly suggestive it will be present in meiotic cells. Our results showing an association between DSB frequencies measured in meiotic cells and enrichment for histone modifications measured in mitotic cells suggests that nucleosome occupancy and H3K4me3 may not be the only chromatin features that mark sites of meiotic DSBs in mitotic cells before the entrance to meiosis. Although, this is a question that cannot be answered by *in silico* analysis because it requires further experimentation measuring the distribution of these marks for both meiotic and mitotic cells.

Role of Histone Modifications

The role of histone modifications in specifying sites of Spo11-catalyzed DSBs is unclear. Specific marks could serve to directly recruit proteins involved in recombination. Alternatively, histone modifications, such as acetylation, may act indirectly by modifying the local chromatin structure. Histone acetyltransferases and ATP-dependent chromatin remodeling factors have been shown to regulate recombination at the *ade6-M26* hotspot in *S. pombe* [15,65]. Deletion of the histone acetyltransferase *GCM5* gene causes a significant delay in chromatin remodeling, leading to a partial reduction in recombination frequency. Double deletion of *SNF22*, a component of a chromatin remodeling complex, and *GCM5* leads to a complete loss of meiotic recombination. *RSC4p*, a component of the chromatin remodeling complex *RSC*, contains tandem bromodomains that recognize H3K14ac, suggesting that this mark may recruit chromatin remodeling factors directly [66]. In addition, acetylation leads to a more open and less condensed chromatin structure, allowing easier access for recombination proteins or chromatin remodeling complexes.

Dot1p the methyltransferase responsible for lysine 79 methylation has been linked with DNA repair [67]. Deletion of Dot1p confers increased sensitivity to radiation in yeast [68]. Additionally the correct function of the DNA checkpoint response requires H3 methylation by Dot1p [69]. The presence of Dot1p is necessary for efficient repair of DSB by sister chromatid repair [70]. This

suggests H3K79me3 may be associated with regions of low meiotic DSBs frequency because it is a marker for DNA repair.

Another possibility is that specific histone modifications may affect DSB frequencies indirectly by inhibiting or enhancing other histone modifications that play a more direct role. For instance, preventing H2B ubiquitination leads to decreased meiotic DSBs[35]. By promoting H3K4me3, H2B ubiquitination may be enhancing DSB formation [71]. Another possible example of similar “cross-talk” between histone modifications is H3K36me3-mediated repression of DSB formation at the well-studied *HIS4* recombination hot spot in budding yeast [44]. H3K36me3 recruits the Rpd3 histone deacetylase [45], suggesting that this mark may have an indirect negative effect on DSB frequency by preventing or reducing histone acetylation since there appears to be a positive correlation between histone acetylation and DSB frequency at some hot spots [15].

Nucleosome mapping

Locke *et al.* [56] were able to predict nucleosome positions using nucleosome free control data they suggest this could be because MNase sequence preference or sonication fragmentation coincides with nucleosome excluding sequence. If this were the case any “peaks” of read density in the MNase or sonication control datasets at hot spots may well reflect true nucleosome occupancy. In support of this hypothesis a recent study in mice found evidence of increased nucleosome binding at hot spots [72].

We do not think this is the case for the genomic loci in question for a number of reasons. One the same set of genomic loci used in our study i.e.(Buhler *et al.* Hot spots) were recently shown to be on average nucleosome depleted using FAIRE [54]. This directly contradicts the sequencing based results at these same loci (Figure 4 panels d, e and f). Two microarray based nucleosome occupancy maps are in agreement with one another but disagree with the results of the uncorrected sequencing based studies (Figure 4). Finally a number of individual hot spots have been examined (see above) and in general they are nucleosome depleted.

The extent to which nucleosome binding is based on sequence preferences is currently an active area of research [23,57]. One approach to answering this question is comparing nucleosome maps produced *in vitro* and *in vivo* [23]. Our results, along with others [73,74], indicate that a systematic bias can dominate at certain genomic loci, thereby obscuring the true biological representation. It is unknown to what extent this influences the genome wide similarity observed between *in vivo* and *in vitro* produced nucleosome occupancy maps.

Using control experiments to remove the systematic bias is an obvious approach in dealing with experimental artifacts. Unfortunately, producing suitable controls is not necessarily straightforward [75]. Previously, controls have rarely been used in nucleosome mapping with high throughput sequencing methods. When experimental bias is not controlled for, the opposite of the most likely correct biological picture is observed at yeast meiotic hot spots mapped at low resolution. However, when we subtract a MNase control experiment from the nucleosome occupancy maps, the correct biological interpretation can be derived indicating the suitability of this control for the loci under investigation in this study. Furthermore, our results underscore the importance of addressing experimental bias in nucleosome mapping high throughput sequencing experiments. Our analysis is not intended to be a comprehensive examination of all possible biological features potentially associated with meiotic DSB frequency Future work could expand the set of genome wide features being examined at sites of high/low meiotic DSB frequencies. Here we have shown feature selection can productively be used to identify

promising biological associations. Our approach successfully identified previously known correlations while making several novel predictions.

Supporting Information

Figure S1 Plots of average modification level around transcription start sites (TSS) using Pan *et al.* hot spots.

Figure is produced as described for Figure 3 with one difference. For panels a, b, d and e genes with a hotspot in their promoter regions were further divided based on the strength of the hot spot. The blue line is the given histone modification plotted upstream of genes whose hot spot is below the first quartile. The red line is genes whose hot spot strength falls between the first and second quartile. The purple line is genes whose hot spots falls between the second and third quartile. The green line is genes whose hot spots strength is greater than the third quartile.

(TIF)

Figure S2 Gene expression comparison in meiotic cells.

Panels a-d is comparing gene expression between genes associated with hot spots to genes not associated with hot spots. Height of bars represents the difference in median gene expression for genes associated with hot spots to genes not associated with hot spots (i.e. Median hot gene expression – Median not hot gene expression). Time points represent time after yeast culture is placed in sporulating media. Panels (a) and (c) represent gene expression measured at the given time points for sporulation deficient SK1 and W303 strains these strains do not enter meiosis. Panels (b) and (d) represent gene expression for sporulation-proficient SK1 and W303 strains. An asterisk represents the difference in medians is significant with $p\text{-value} < 0.05$, $p\text{-value}$ calculated using the Wilcoxon rank sum test. Panels e-h is as described above except height of bars represents the difference in median gene expression for genes associated with cold spots to genes not associated with cold spots (i.e. Median cold gene expression – Median not cold gene expression). Gene expression is represented by hybridization fluorescence intensities.

(TIF)

Figure S3 Nucleosome occupancy at Buhler *et al.* hot spots for all sequencing-based datasets.

For all datasets, reads were mapped to the yeast genome. Only uniquely aligned reads were retained and the count mapped to each base pair was calculated. The z-score standardized count of reads is plotted using the same procedure as described for Figure 4 with the green line representing GC content. (references for datasets: a [76], b [23], c [77], d [78], e [63] and f [79].

(TIF)

Figure S4 Read density for sequencing controls at Pugh *et al.* hot spots.

(a) Purified DNA digested with micrococcal nuclease (MNase) and sequenced using the Solexa platform. (b) Purified DNA following sonication and sequencing using the Solexa platform. The black line indicates the z-score standardized mapped read density. Data was smoothed using loess smoothing.

(TIF)

References

1. Lichten M, Goldman AS (1995) Meiotic recombination hotspots. Annual Review of Genetics 29: 423–444.
2. Martinez-Perez E, Colaiacovo MP (2009) Distribution of meiotic recombination events: talking to your neighbors. Current Opinion in Genetics & Development 19: 105–112.
3. Keeney S, Giroux CN, Kleckner N (1997) Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. Cell 88: 375–384.

Figure S5 Nucleosome occupancy at Pugh *et al.* hot spots for all sequencing-based datasets.

For all datasets, reads were mapped to the yeast genome. Only uniquely aligned reads were retained and the count mapped to each base pair was calculated. The z-score standardized count of reads is plotted at centered Pugh *et al.* hot spots. Plot is produced similar to Figure 4 and Figure S3. (References for datasets: a [76], b [23], c [77], d [78], e [63] and f [79].

(TIF)

Figure S6 Nucleosome occupancy at recombination hot spots obtained at various resolutions.

Z-score standardized nucleosome occupancy is shown in 100 bp bins (y-axis). The center of the aligned hot spots is zero on the x-axis. Panels a, b and c represent nucleosome occupancy data measured by ChIP-chip produced by Lee *et al.* [55] at three different hot spot datasets from left to right [5], [14], and [21]. Panels d, e and f represent nucleosome occupancy in the same three datasets but now using a nucleosome occupancy map produced by ChIP-seq [64]. This sequencing based nucleosome occupancy map has previously been used in analyzing nucleosome occupancy at hot spots as defined by Borde *et al.* [14].

(TIF)

Figure S7 Sonication and MNase control plotted at Buhler *et al.* hot spots allowing multimapping reads.

Reads for sonicated (a) and MNase-digested controls (b) were mapped allowing multimapping of reads. Read density centered at hot spots is plotted. Data was smoothed using loess smoothing.

(TIF)

Figure S8 Nucleosome occupancy at Buhler *et al.* hot spots for all sequencing-based datasets following subtraction of the MNase control.

Nucleosome occupancy was plotted at hot spots for all sequencing-based nucleosome mapping datasets following subtraction of the MNase control as described in the text. Data plotted similarly to Figure 4.

(TIF)

Acknowledgments

We are grateful to Michael Lichten for advice and for critically reading this manuscript. We are also grateful to Jean and Danielle Thierry-Mieg for multiple discussions on the results and interpretation of this work and for their constructive suggestions. We would also like to thank Istvan Albert who provided data and advice. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (<http://biowulf.nih.gov>).

Author Contributions

Conceived and designed the experiments: LH DL LM-R. Performed the experiments: LH N-KK. Analyzed the data: LH N-KK. Contributed reagents/materials/analysis tools: LH N-KK. Wrote the paper: LH DL LM-R.

7. Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, et al. (2000) Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences, USA* 97: 11383–11390.
8. Bagshaw AT, Pitt JP, Gemmell NJ (2008) High frequency of microsatellites in *S. cerevisiae* meiotic recombination hotspots. *BMC Genomics* 9: 49.
9. Zhou T, Weng J, Sun X, Lu Z (2006) Support vector machine for classification of meiotic recombination hotspots and coldspots in *Saccharomyces cerevisiae* based on codon composition. *BMC Bioinformatics* 7: 223.
10. Jiang P, Wu H, Wei J, Sang F, Sun X, et al. (2007) RF-DYMHMC: detecting the yeast meiotic recombination hotspots and coldspots by random forest model using gapped dinucleotide composition features. *Nucleic Acids Research* 35: W47–51.
11. Wu TC, Lichten M (1994) Meiosis-induced double-strand break sites determined by yeast chromatin structure. *Science* 263: 515–518.
12. Ohta K, Shibata T, Nicolas A (1994) Changes in chromatin structure at recombination initiation sites during yeast meiosis. *EMBO Journal* 13: 5754–5763.
13. Murakami H, Borde V, Shibata T, Lichten M, Ohta K (2003) Correlation between premeiotic DNA replication and chromatin transition at yeast recombination initiation sites. *Nucleic Acids Research* 31: 4085–4090.
14. Borde V, Robine N, Lin W, Bonfils S, Geli V, et al. (2009) Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *EMBO Journal* 28: 99–111.
15. Yamada T, Mizuno K, Hirota K, Kon N, Wahls WP, et al. (2004) Roles of histone acetylation and chromatin remodeling factor in a meiotic recombination hotspot. *EMBO Journal* 23: 1792–1803.
16. Ooi CH, Tan P (2003) Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics* 19: 37–44.
17. Sha N, Vannucci M, Tadesse MG, Brown PJ, Dragoni I, et al. (2004) Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* 60: 812–819.
18. Xiong M, Fang X, Zhao J (2001) Biomarker identification by feature wrappers. *Genome Research* 11: 1878–1887.
19. Shin H, Sheu B, Joseph M, Markey MK (2008) Guilt-by-association feature selection: identifying biomarkers from proteomic profiles. *Journal of Biomedical Informatics* 41: 124–136.
20. Saey Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* 23: 2507–2517.
21. Pan J, Sasaki M, Kniwel R, Murakami H, Blitzzblau HG, et al. (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* 144: 719–731.
22. Song JS, Johnson WE, Zhu X, Zhang X, Li W, et al. (2007) Model-based analysis of two-color arrays (MA2C). *Genome Biology* 8: R178.
23. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, et al. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458: 362–366.
24. White MA, Dominska M, Petes TD (1993) Transcription factors are required for the meiotic recombination hotspot at the *HIS4* locus in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences, USA* 90: 6621–6625.
25. David EG (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*: Addison-Wesley Longman Publishing Co., Inc. 372 p.
26. Hansen L, Lee EA, Hestir K, Williams LT, Farrelly D (2009) Controlling feature selection in random forests of decision trees using a genetic algorithm: classification of class I MHC peptides. *Combinatorial chemistry & high throughput screening* 12: 514–519.
27. Li L, Darden TA, Weinberg CR, Levine AJ, Pedersen LG (2001) Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Combinatorial chemistry & high throughput screening* 4: 727–739.
28. Li L, Weinberg CR, Darden TA, Pedersen LG (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* 17: 1131–1142.
29. Trevino V, Falciani F (2006) GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics* 22: 1154–1156.
30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.
31. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, et al. (2007) Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* 446: 572–576.
32. Xu H, Wei CL, Lin F, Sung WK (2008) An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics* 24: 2344–2349.
33. Sollier J, Lin W, Soustelle C, Suhre K, Nicolas A, et al. (2004) Set1 is required for meiotic S-phase onset, double-strand break formation and middle gene expression. *EMBO Journal* 23: 1957–1967.
34. Sun ZW, Allis CD (2002) Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. *Nature* 418: 104–108.
35. Yamashita K, Shinohara M, Shinohara A (2004) Rad6-Bre1-mediated histone H2B ubiquitylation modulates the formation of double-strand breaks during meiosis. *Proceedings of the National Academy of Sciences, USA* 101: 11380–11385.
36. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, et al. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327: 876–879.
37. Parvanov ED, Petkov PM, Paigen K (2010) Prdm9 controls activation of mammalian recombination hotspots. *Science* 327: 835.
38. Baudat F, Buard J, Grey C, Fedel-Alon A, Ober C, et al. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327: 836–840.
39. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, et al. (2005) Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biology* 3: e328.
40. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 122: 517–527.
41. Rando OJ (2007) Global patterns of histone modifications. *Current Opinion in Genetics & Development* 17: 94–99.
42. Blander G, Guarente L (2004) The Sir2 family of protein deacetylases. *Annual Review of Biochemistry* 73: 417–435.
43. Mieczkowski PA, Dominska M, Buck MJ, Lieb JD, Petes TD (2007) Loss of a histone deacetylase dramatically alters the genomic distribution of Spo11p-catalyzed DNA breaks in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences, USA* 104: 3955–3960.
44. Merker JD, Dominska M, Greenwell PW, Rinella E, Bouck DC, et al. (2008) The histone methylase Set2p and the histone deacetylase Rpd3p repress meiotic recombination at the *HIS4* meiotic recombination hotspot in *Saccharomyces cerevisiae*. *DNA Repair (Amst)* 7: 1298–1308.
45. Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, et al. (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* 123: 581–592.
46. Kolasinska-Zwiercz P, Down T, Latorre I, Liu T, Liu XS, et al. (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genetics* 41: 376–381.
47. Gat-Viks I, Vingron M (2009) Evidence for gene-specific rather than transcription rate-dependent histone H3 exchange in yeast coding regions. *PLoS Computational Biology* 5: e1000282.
48. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95: 717–728.
49. Primig M, Williams RM, Winzler EA, Tevzadze GG, Conway AR, et al. (2000) The core meiotic transcriptome in budding yeasts. *Nature Genetics* 26: 415–423.
50. Baudat F, Nicolas A (1997) Clustering of meiotic double-strand breaks on yeast chromosome III. *Proceedings of the National Academy of Sciences, USA* 94: 5213–5218.
51. Liu J, Wu TC, Lichten M (1995) The location and structure of double-strand DNA breaks induced during yeast meiosis: evidence for a covalently linked DNA-protein intermediate. *EMBO Journal* 14: 4599–4608.
52. Mizuno K, Emura Y, Baur M, Kohli J, Ohta K, et al. (1997) The meiotic recombination hot spot created by the single-base substitution ade6M26 results in remodeling of chromatin structure in fission yeast. *Genes & Development* 11: 876–886.
53. Hirota K, Steiner WW, Shibata T, Ohta K (2007) Multiple modes of chromatin configuration at natural meiotic recombination hot spots in fission yeast. *Eukaryotic Cell* 6: 2072–2080.
54. Berchowitz LE, Hanlon SE, Lieb JD, Copenhaver GP (2009) A positive but complex association between meiotic double-strand break hotspots and open chromatin in *Saccharomyces cerevisiae*. *Genome Research* 19: 2245–2257.
55. Lee W, Tillo D, Bray N, Morse RH, Davis RW, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics* 39: 1235–1244.
56. Locke G, Tolkmunov D, Moqtaderi Z, Struhl K, Morozov AV (2010) High-throughput sequencing reveals a simple model of nucleosome energetics. *Proceedings of the National Academy of Sciences, USA* 107: 20998–21003.
57. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, et al. (2009) Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nature Structural & Molecular Biology* 16: 847–852.
58. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* 36: e105.
59. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, et al. (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods* 5: 183–188.
60. Fan HC, Quake SR (2010) Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. *PLoS One* 5: e10439.
61. Cheung MS, Down TA, Latorre I, Ahinger J (2011) Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Research*.
62. Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL (2004) Global nucleosome occupancy in yeast. *Genome Biology* 5: R62.
63. Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, et al. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research* 18: 1073–1083.
64. Zhang L, Ma H, Pugh BF (2011) Stable and dynamic nucleosome states during a meiotic developmental process. *Genome Research* 21: 875–884.

65. Hirota K, Mizuno K, Shibata T, Ohta K (2008) Distinct chromatin modulators regulate the formation of accessible and repressive chromatin at the fission yeast recombination hotspot *ade6-M26*. *Molecular Biology of the Cell* 19: 1162–1173.
66. Kasten M, Szerlong H, Erdjument-Bromage H, Tempst P, Werner M, et al. (2004) Tandem bromodomains in the chromatin remodeler RSC recognize acetylated histone H3 Lys14. *EMBO Journal* 23: 1348–1359.
67. van Leeuwen F, Gafken PR, Gottschling DE (2002) Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell* 109: 745–756.
68. Game JC, Williamson MS, Baccari C (2005) X-ray survival characteristics and genetic analysis for nine *Saccharomyces* deletion mutants that show altered radiation sensitivity. *Genetics* 169: 51–63.
69. Giannattasio M, Lazzaro F, Plevani P, Muzi-Falconi M (2005) The DNA damage checkpoint response requires histone H2B ubiquitination by Rad6-Bre1 and H3 methylation by Dot1. *Journal of Biological Chemistry* 280: 9879–9886.
70. Conde F, Refolio E, Cordon-Preciado V, Cortes-Ledesma F, Aragon L, et al. (2009) The Dot1 histone methyltransferase and the Rad9 checkpoint adaptor contribute to cohesin-dependent double-strand break repair by sister chromatid recombination in *Saccharomyces cerevisiae*. *Genetics* 182: 437–446.
71. Kniewel R, Keeney S (2009) Histone methylation sets the stage for meiotic DNA breaks. *EMBO Journal* 28: 81–83.
72. Smagulova F, Gregoret IV, Brick K, Khil P, Camerini-Otero RD, et al. (2011) Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature*.
73. Fan X, Moqtaderi Z, Jin Y, Zhang Y, Liu XS, et al. (2010) Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation. *Proceedings of the National Academy of Sciences, USA* 107: 17945–17950.
74. Chung HR, Dunkel I, Heise F, Linke C, Krobisch S, et al. (2010) The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One* 5: e15754.
75. Kaplan N, Hughes TR, Lieb JD, Widom J, Segal E (2010) Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. *Genome Biology* 11: 140.
76. Field Y, Kaplan N, Fondulke-Mittendorf Y, Moore IK, Sharon E, et al. (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Computational Biology* 4: e1000216.
77. Eaton ML, Galani K, Kang S, Bell SP, MacAlpine DM (2010) Conserved nucleosome positioning defines replication origins. *Genes & Development* 24: 748–753.
78. Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N (2010) High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Research* 20: 90–100.
79. Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ (2010) The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biology* 8: e1000414.

Transcriptional Regulation of N-Acetylglutamate Synthase

Sandra Kirsch Heibel^{1,2}, Giselle Yvette Lopez³, Maria Panglao⁴, Sonal Sodha⁵, Leonardo Mariño-Ramírez⁶, Mendel Tuchman¹, Ljubica Caldovic^{1*}

1 Center for Genetic Medicine Research, Children's National Medical Center, Washington, D. C., United States of America, **2** Molecular and Cellular Biology Program, University of Maryland, College Park, Maryland, United States of America, **3** Department of Pathology, Duke University Medical Center, Durham, North Carolina, United States of America, **4** The George Washington University School of Medicine and Health Sciences, Washington, D. C., United States of America, **5** Johns Hopkins School of Medicine in Baltimore, Maryland, United States of America, **6** Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America

Abstract

The urea cycle converts toxic ammonia to urea within the liver of mammals. At least 6 enzymes are required for ureagenesis, which correlates with dietary protein intake. The transcription of urea cycle genes is, at least in part, regulated by glucocorticoid and glucagon hormone signaling pathways. N-acetylglutamate synthase (NAGS) produces a unique cofactor, N-acetylglutamate (NAG), that is essential for the catalytic function of the first and rate-limiting enzyme of ureagenesis, carbamyl phosphate synthetase 1 (CPS1). However, despite the important role of NAGS in ammonia removal, little is known about the mechanisms of its regulation. We identified two regions of high conservation upstream of the translation start of the *NAGS* gene. Reporter assays confirmed that these regions represent promoter and enhancer and that the enhancer is tissue specific. Within the promoter, we identified multiple transcription start sites that differed between liver and small intestine. Several transcription factor binding motifs were conserved within the promoter and enhancer regions while a TATA-box motif was absent. DNA-protein pull-down assays and chromatin immunoprecipitation confirmed binding of Sp1 and CREB, but not C/EBP in the promoter and HNF-1 and NF-Y, but not SMAD3 or AP-2 in the enhancer. The functional importance of these motifs was demonstrated by decreased transcription of reporter constructs following mutagenesis of each motif. The presented data strongly suggest that Sp1, CREB, HNF-1, and NF-Y, that are known to be responsive to hormones and diet, regulate *NAGS* transcription. This provides molecular mechanism of regulation of ureagenesis in response to hormonal and dietary changes.

Citation: Heibel SK, Lopez GY, Panglao M, Sodha S, Mariño-Ramírez L, et al. (2012) Transcriptional Regulation of N-Acetylglutamate Synthase. PLoS ONE 7(2): e29527. doi:10.1371/journal.pone.0029527

Editor: Venugopalan Cheriyaath, Texas A&M University, United States of America

Received: February 24, 2011; **Accepted:** November 30, 2011; **Published:** February 27, 2012

This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Funding: This work was supported by Public Health Service Grant R01DK064913 from the National Institutes of Health. This research was also supported in part by the Intramural Research Program of the NIH, NLM, NCBI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ljubica@cnmcresearch.org

Introduction

Ammonia, the toxic product of protein catabolism, is converted to urea by the urea cycle in the liver of mammals. Incorporation of two nitrogen atoms into urea is catalyzed by six enzymes: three of them mitochondrial, N-acetylglutamate synthase (NAGS; EC 2.3.1.1), carbamylphosphate synthetase 1 (CPS1; EC 6.4.3.16) and ornithine transcarbamylase (OTC; EC 2.1.3.3), and the other three cytosolic, argininosuccinate synthetase (ASS; EC 6.3.4.5), argininosuccinate lyase (ASL; EC 4.3.2.1) and arginase 1 (Arg1; EC 3.5.3.1).

NAGS catalyzes the formation of N-acetylglutamate (NAG), an essential allosteric activator of CPS1, in the mitochondrial matrix of hepatocytes and small intestine epithelial cells [1,2]. Within hepatocytes, NAGS activity and NAG abundance are regulated by L-arginine, ammonia, and dietary protein intake [3,4,5] and therefore, the NAGS/NAG system may play a critical role in the regulation of ureagenesis in response to these factors [6]. While studies in the 1980s and 1990s identified the *cis*-acting motifs regulating transcription of the urea cycle enzymes CPS1

[7,8,9,10], OTC [11,12,13,14], ASS [15,16,17], ASL [18,19,20], and Arg1 [21,22], the mammalian *NAGS* gene was not identified until 2002 [2] and we can now report for the first time on its transcriptional regulation.

Many studies have identified regulatory links between the urea cycle genes and glucocorticoids and glucagon [23,24,25], however the mechanism of regulation differs for each gene [24,26,27,28,29]. Transcription of *CPS1* is activated by TATA-binding protein (TBP) while its proximal and distal enhancers contain binding sites for glucocorticoids and cAMP responsive factors including CCAAT-enhancer bind protein (C/EBP), activator protein-1 (AP-1), glucocorticoid receptor (GR) and cAMP response element binding (CREB). Sites for binding tissue specific factors including hepatic nuclear factor 3 (HNF-3) are also present [25,30,31]. Tissue specific expression of the *OTC* gene is induced in the intestine and liver by HNF-4, which binds in the promoter [13,14,32] while binding of both HNF-4 and C/EBP to the enhancer, induces high expression levels in the liver [12,13,14,25,33]. *ASS* transcription is regulated by cooperative binding of multiple specificity protein 1 (Sp1) [16,34,35,36]. *ASL* is regulated through Sp1 and the positive

regulator, nuclear factor Y (NF-Y), which binds within the promoter of *ASL* to activate its transcription [18,19,20,37]. Sp1 and nuclear factor 1 (NF-1)/CCAAT-binding transcription factor (CTF) activate *ARG1* transcription while two C/EBP factors and two unidentified proteins bind within an enhancer in intron 7 to confer glucocorticoid responsiveness [22].

Abundance of urea cycle enzymes correlates with dietary protein intake [3,28]. Transcription of urea cycle genes is in part regulated by the glucocorticoid and glucagon signaling pathways [29,38]. Therefore, we postulate that there exists a nitrogen sensing mechanism that is both responsive to amino acid(s) and hormone stimulation and that an understanding of the transcriptional regulation of *NAGS* could contribute to the understanding of such mechanism.

In this study, we identified two regulatory regions upstream of the *NAGS* translation start site that contain highly conserved protein-binding DNA motifs. We subsequently confirmed that these regions function as promoter and enhancer and that the enhancer is most effective in liver cells. Avidin-agarose protein-DNA pull-down assays have been used to confirm binding of Sp1 and CREB within the *NAGS* promoter and Hepatic Nuclear Factor 1 (HNF-1) and NF-Y within the enhancer regions. Chromatin immunoprecipitation (ChIP) and quantitative real-time PCR have been used to independently verify that Sp1 and CREB bind to the promoter region, and HNF-1 and NF-Y bind to the enhancer region. We also used 5'RACE analysis to identify multiple transcription start sites for *NAGS* that may be species and tissue specific. These findings provide new information on the regulation of the *NAGS* gene, and suggest possible mechanisms for coordinated regulation of the genes involved in ureagenesis.

Materials and Methods

Bioinformatic Analysis of the Upstream Regulatory Regions

Pair-wise Alignment Analysis. Identification of highly conserved regions was conducted by gathering 15 kilobases of genomic sequence 5' of the *NAGS* translational start site and sequence of intron one in 7 mammalian species including: human (NM_153006.2), chimpanzee (XM_001152480.1), dog (XM_548066.2), cow (XM_618194.4), horse (XM_001917302.1), mouse (NM_145829.1) and rat (NM_001107053.1). The highly conserved regulatory regions of *CPS1* were identified by gathering 15 kilobases of genomic sequences 5' of the translational start site from human (NM_001875), chimpanzee (XM_001146604), dog (XM_856862), mouse (NM_001080809), and rat (NM_017072). Genomic sequences were subject to pair-wise comparison using BLAST bl2seq tool [39]. Parameters included expect threshold of 10, match and mismatch scores of 1 and -2, respectively, gap existence and extension scores of 5 and 2 respectively, and maximum expected value $E=0.001$. Regions of high conservation were identified as sequences with more than 80% identity that were at least 100 bp long and present in four or more species.

Cis-element OVERrepresentation (CLOVER) Analysis. The Cis-element OVERrepresentation (CLOVER) [40] program was used to predict the over-represented motifs within the highly conserved regulatory regions of *NAGS* and *CPS1*. CLOVER analysis of these conserved regions identified known protein binding DNA motifs in the TRANSFAC Pro database by calculating over-representation of these sequences compared to a background of ppr_build_33.fa generated from NCBI build 33 [41]. Matrices recognized by multiple transcription factors in the same family are represented by one family member unless otherwise noted. Genomic sequences of the highly conserved regions were aligned using CLUSTALW version 2.0.10 [42].

Plasmid Constructs

The promoter and enhancer of *NAGS*, were amplified from human genomic DNA with primer pairs hPromXH and hEnhXH or hPromHXrev and hEnhHXrev (Table S1), respectively, to introduce *XhoI* and *HindIII* restriction enzyme sites and allow subcloning in forward and reverse orientation. Platinum Taq PCRx DNA Polymerase (Invitrogen) was used for amplification with the following conditions: initial denaturation at 95°C for 2 min., followed by 35 cycles of denaturation at 95°C for 30 sec., annealing at 57°C for 30 sec. and extension at 68°C for 1 min., and final extension at 68°C for 6 min. Promoter and enhancer PCR products were ligated with TOPO-TA sequencing vector (Invitrogen) according to manufacturer's instructions and referred to as TOPOProm, TOPOEnh, TOPOPromRev, and TOPOEnhRev, respectively. Mouse *Nags* (mNags) promoter and enhancer were inserted into TOPO-TA vector following the same methods. Correct DNA sequences were confirmed using sequencing primers specified by Invitrogen.

TOPOProm, TOPOEnh, TOPOPromRev, TOPOEnhRev, pGL4.10 (Promega) basic vector containing firefly (*Photinus pyralis*) luciferase *luc2*, and pGL4.23 (Promega) vector containing a minimal TATA promoter with *luc2* were cut with *XhoI* (New England Biolabs) and *HindIII* (New England Biolabs). The vectors were treated with Antarctic Alkaline Phosphatase (AAP) (New England Biolabs) according to manufacturer's instructions, and the *NAGS* regions were ligated with the vectors to form the plasmids in Table 1. TOPOEnh was also amplified with primer pair hEnhBS (Table S1), to introduce *BamHI* and *Sall* restriction enzyme sites at the 5' and 3' ends of the enhancer, respectively. The amplified enhancer product and 4.10Prom were cut with *BamHI* (New England Biolabs) and *Sall* (New England Biolabs), the vector was treated with AAP, and the enhancer was ligated with the vector (Table 1). Plasmids containing mouse *NAGS* promoter and enhancer were generated using the same methods with the primer pairs listed in Table S1 and plasmids in Table 1. Correct sequences were confirmed using primers specified by Promega.

Point mutations in the binding sites for transcription factors Sp1, HNF-1 and NF-Y were selected based on functional analysis

Table 1. Plasmids generated for luciferase reporter assays.

Name	Vector	Insert
4.10Prom	pGL4.10	hNAGS promoter
4.10Enh	pGL4.10	hNAGS enhancer
4.23Enh	pGL4.23	hNAGS enhancer
4.10PromEnh	4.10Prom	hNAGS enhancer
4.10PromRev	pGL4.10	hNAGS promoter reverse
4.23EnhRev	pGL4.23	hNAGS enhancer reverse
m4.10Prom	pGL4.10	mNAGS promoter
m4.10Enh	pGL4.10	mNAGS enhancer
m4.23Enh	pGL4.23	mNAGS enhancer
m4.10PromEnh	4.10Prom	mNAGS enhancer
4.10Sp1m	pGL4.10	hNAGS promoter with Sp1 mutations
4.10CREBm	pGL4.10	hNAGS promoter with CREB mutations
4.23HNF-1m	pGL4.23	hNAGS enhancer with HNF-1 mutations
4.23NF-Ym	pGL4.23	hNAGS enhancer with NF-Y mutations

Human or mouse promoter or enhancer were ligated with pGL4 vectors for use with luciferase reporter assays.

doi:10.1371/journal.pone.0029527.t001

of Sp1 [43,44,45], HNF-1 [46,47], and NF-Y [48,49] binding in other genes. Mutations were engineered by Integrated DNA Technologies and provided in pIDTSMART-KAN vectors (IDT) (Table 2). Plasmids with mutant Sp1, HNF-1, and NFY were cut with *XhoI* and *HindIII*. Reporter plasmids pGL4.10, and pGL4.23 were cut with *XhoI* and *HindIII* and treated with AAP. Mutated inserts were ligated with vectors to form the plasmids 4.10Sp1m, 4.23HNF-1m, and 4.23NFYm (Table 1). Correct sequences were confirmed using primers specified by Promega.

Point mutations in the CREB binding site, c.-7T>C and c.-5T>A (Table 2), were selected based on functional analysis of CREB binding [50,51] in other genes and were engineered into the *NAGS* gene using QuickChange Lightning Site-Directed Mutagenesis Kit (Agilent) according to manufacturer's instructions. Primers hCREBm Fw and Rv (Table S1) amplified 50 ng of template plasmid 4.10Prom to create 4.10CREBm. The correct sequence was confirmed using primers specified by Promega.

The expression vectors encoding Sp1 or HNF-1 cDNA were under control of the cytomegalovirus promoter (Origene).

Tissue culture

Cell culture and transfection. Human hepatoma cells (HepG2) (donated by Dr. Marshall Summar, Children's National Medical Center, Washington, DC) were cultured in complete media containing RPMI 1640 medium (Invitrogen) supplemented with 10% fetal bovine serum (FBS) (ATCC) and 5% Penicillin/Streptomycin (Invitrogen) under 5% CO₂ at 37°C. Human alveolar basal epithelial cells (A549) (donated by Dr. Mary Rose, Children's National Medical Center, Washington, DC) were cultured in complete media containing Ham's F-12 medium (Invitrogen) supplemented with 10% FBS and 5% Penicillin/Streptomycin. Human colorectal adenocarcinoma cells (Caco-2) (ATCC) were cultured in Eagle's Minimum Essential Medium (Invitrogen) supplemented with 20% FBS. Cells were plated at a density of 5×10^5 cells/well on 24-well culture plates 24 hours prior to transfection. The cells (90–95% confluent for HepG2 and A549, 80–85% confluent for Caco-2) were then transfected using Lipofectamine 2000 reagent (Invitrogen) and cultured in transfection media containing medium and serum only. A total of 0.25 µg of DNA was transfected with 0.225 µg of vector expressing *luc2* and 0.025 µg of pGL4.74 vector containing *Renilla reniformis* luciferase (*hRluc*) as an internal control (Promega). For co-transfections 0.225 µg of *luc2* vector was combined with either 0.25 µg of expression vector or empty vector pUC19 (Invitrogen), and 0.025 µg of *hRluc* control vector.

Reporter assays

24 hours following transfection, cells were assayed for both firefly and *Renilla* luciferase activity using Dual-Luciferase Reporter Assay System (Promega) and Berthold Centro 960 luminometer (Berthold) according to the manufacturer's protocol. All reporter assay measurements were corrected for transfection efficiency by normalizing the firefly luciferase signal to the *Renilla* luciferase values. Expression level of each construct was determined relative to luciferase expression under control of the *NAGS* promoter in each cell line. All results are an average of three independent experiments that were each carried out in triplicate. Values were expressed as mean \pm SEM and analyzed using Student's *t*-test.

5' Rapid Amplification of cDNA Ends (RACE)

5' RACE (Version 2.0; Invitrogen) was performed using RNA isolated from donated mouse livers by Trizol reagent (Invitrogen). RNA from mouse small intestine (Origene), human duodenum (Ambion), or human liver (Ambion) was commercially available. Products were synthesized with human or mouse *NAGS* specific primers complementary to sequence within Exon 1 (Table S2). All reactions began with 5 µg of total RNA and the RACE procedure was conducted according to manufacturer's instructions. Second strand synthesis was conducted using Ex Taq Polymerase (TaKaRa Bio Inc.). PCR products were subcloned into pCR 2.1-TOPO vector (Invitrogen) and RACE products were sequenced with primers specified by the manufacturer.

Avidin-Agarose DNA-Protein Pull-Down Assay

Biotinylated DNA probes. Probes for Avidin-Agarose DNA-Protein Pull-Down Assays were generated by PCR amplification of genomic DNA isolated from donated mouse tails using Pure Gene DNA Purification Kit (Gentra). Probes were generated using biotinylated or non-biotinylated forward primer and non-biotinylated reverse primers with Platinum Taq PCRx DNA Polymerase (Invitrogen) and amplification conditions: initial denaturation at 95°C for 2 min., followed by 35 cycles of denaturation at 95°C for 30 sec., annealing at 60°C for 30 sec. and extension at 68°C for 1 min., and final extension at 68°C for 6 min. The mouse *Nags* (*mNags*) promoter regions A and B (Figure 1) were amplified with primer pair mNAGS-Prom Region A, from +97 to −259, relative to the translation initiation codon and with mNAGS-Prom Region B, from −302 to −776, respectively (Table S3). A region of *mNags*, that is not highly conserved in mammals, −1056 to −1320, was amplified using primer pair mNAGS-Prom-NC to serve as a negative control for the promoter regulatory region. The enhancer region of mNAGS, spanning from −2834 to −3167, was amplified using forward primer pair mNAGS-enh. The negative control for the enhancer region, a non-conserved region located close to enhancer, was the amplification product of primer pair mNAGS-Enh-NC spanning −5569 to −5997 upstream of *mNags*. Additional negative controls, non-biotinylated probes, were generated using each primer pair.

Preparation of nuclear extracts. Nuclear extract was isolated from donated adult mouse livers of C57BL/6 mice using Nuclear Extraction Kit (Origene) according to manufacturer's instructions. The protein concentration of the nuclear extract was determined using bovine serum albumin as the protein standard with Bradford Assay dye concentrate reagents (Bio-Rad). On average, 10 mg of nuclear protein was obtained from mouse liver.

Binding Protocol and Western Blot. For the avidin-agarose protein-DNA pull-down assay [52], 1 mg of nuclear extract in PBS buffer containing inhibitors (PBSI; 1× PBS with 0.5 mM PMSF,

Table 2. Mutations in Sp1 and CREB binding sites in the promoter, and HNF-1 and NF-Y in the enhancer of human *NAGS*.

Factor	Wild-type	Mutant
Sp1	5'-CCGCCCCCGCC-3'	5'-AAGAACAAGAA-3'
	5'-GGGGCGGGGG-3'	5'-GGTCTTTGG-3'
	5'-CCCCGCCCCC-3'	5'-CCAAGAAACC-3'
	5'-CCCCGCCCGC-3'	5'-CCAAGAAACG-3'
CREB	5'-GGTGTGTCGTCATGG-3'	5'-GGTCGACGTCATGG-3'
HNF-1	5'-TGGAGTTAATCATCTACTCTG-3'	5'-TGGAGTAAGTCTGAACCAAGG-3'
NF-Y	5'-GGCCCCATTGGCTGCCT-3'	5'-GGCCCCCTCCAGCTG-3'

doi:10.1371/journal.pone.0029527.t002

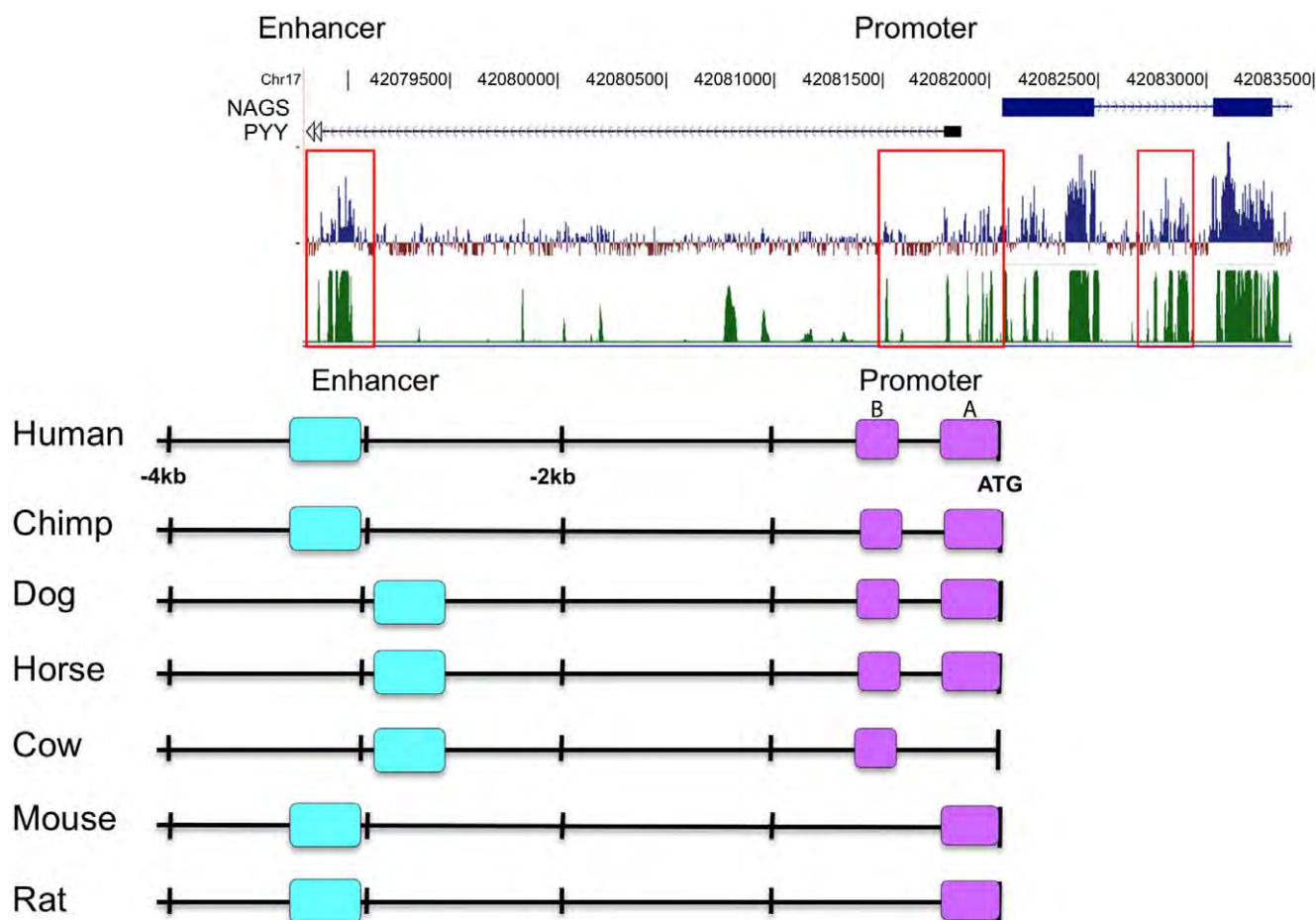


Figure 1. Regions upstream of the mammalian *NAGS* genes that are highly conserved. Conservation of mammalian *NAGS* DNA by phastCons (green) and phyloP (blue) algorithms is shown with the highly-conserved regions indicated in red boxes (A). Pair-wise blast analysis of mammalian non-coding regions of *NAGS* identified highly conserved sequences upstream of the translational start site termed the promoter (purple) and enhancer (cyan) (B).

doi:10.1371/journal.pone.0029527.g001

25 mM β -glycerophosphate, mM NaF), 15 μ g of DNA probe, and avidin-agarose beads (Sigma) were combined and incubated for 16 hrs on a rotating shaker at 4°. The probe and bead concentrations were in excess to ensure complete pull-down of DNA-protein complexes. Following incubation, the supernatant was reserved while the beads were washed 3 times with cold PBSI and then resuspended and boiled in Laemmli protein denaturing buffer (Bio-Rad) with 0.2 M DTT. The supernatant was also combined with denaturing buffer with DTT and boiled; all samples were loaded onto 10% SDS-polyacrylamide gel. The proteins were separated by electrophoresis, transferred to a nitrocellulose membrane, and then identified by immunoblotting using primary antibodies at 1:2000 dilution of antibody to Sp1 (Santa Cruz Biotech; Millipore), 1:1000 dilution of CREB-1 α / β (Santa Cruz Biotech), and 1:3000 dilution of C/EBP α / β (Santa Cruz Biotech) for the promoter region and 1:500 dilution of HNF-1 α / β (Santa Cruz Biotech), 1:1000 dilution of NF-Y α (Santa Cruz Biotech) and 1:2000 dilution of SMAD2/3 (Santa Cruz Biotech) for the -3 kb conserved region. The membrane was then incubated with 1:20,000 dilution of donkey anti-rabbit secondary antibody conjugated to horseradish peroxidase (Pierce) and bands were visualized using SuperSignal West Pico Kit (Pierce) according to manufacturer's instructions.

Chromatin Immunoprecipitation

Tissue preparation and DNA immunoprecipitation. Donated livers from adult C57BL/6 mice were minced and chromatin was precipitated using SimpleChIP Enzyme Chromatin Kit (Origene) with the variation for whole tissue. Briefly, fresh tissue was minced and washed with PBS including Protease Inhibitor Complete tablets (Roche). Proteins and DNA were cross-linked with 1.5% formaldehyde, and tissue was disaggregated with dounce homogenizer. Chromatin was sheared to an approximate size of 100–1000 bp by micrococcal nuclease digestion followed by sonication. Immunoprecipitation was conducted using antibodies to transcription factors Sp1 (Millipore), CREB (Santa Cruz Biotech), C/EBP (Santa Cruz Biotech), HNF-1 (Santa Cruz Biotech), NF-Y (Santa Cruz Biotech), SMAD2 (Santa Cruz Biotech) and AP-2 (Santa Cruz Biotech) and control antibodies to histone H3 and non-specific rabbit IgG (Cell Signaling Technologies). Chromatin was eluted from protein G agarose beads, cross-linking was reversed, and DNA was purified according to manufacturer's instructions.

Real-time PCR quantification. ChIP enriched DNA samples included 2% input control and dilutions for a standard curve, positive control immunoprecipitate from anti-histone H3 antibody sample, negative control immunoprecipitation from

anti-rabbit IgG antibody, no antibody control, water control, and test antibodies. Enriched DNA was subject to quantitative real-time PCR using iTaq SYBR Green Supermix with ROX (Bio-Rad) and gene specific primers (Table S4) including negative locus primers to Chemokine ligand 2 (MIP-2) on a 7900HT Fast Real-Time PCR System (Applied Biosystems). Amplification conditions included initial denaturation at 95°C for 2 min., followed by 50 cycles of denaturation at 95°C for 30 sec., annealing at 60°C for 30 sec. and extension at 72°C for 30 sec., with dissociation steps of 95° for 15 sec. followed by 50° for 15 sec. and finally 95° for 15 sec. Samples were amplified and analyzed using 7900HT Sequence Detection System Software (Applied Biosystems). Values were expressed as mean \pm SEM and analyzed using Student's *t*-test.

Results

Selected regions of non-coding DNA upstream of NAGS are highly conserved

15 kilobase of genomic DNA sequence 5' of the translational start site of *NAGS* and sequence of the first intron from human, chimpanzee, dog, horse, cow, mouse and rat were aligned and compared using pair-wise BLAST. Comparisons showed three highly conserved regions upstream of human *NAGS* at −57 to −284, −498 to −576, and −2978 to −3344 relative to the start ATG, and no significant conservation within the intron or between −5 and −15 kb upstream (Figure 1). The region within −1 kb of the translational start site was designated as the putative promoter while the region 3 kb upstream was designated a putative regulatory element. Figure 1 also shows an alignment of mammalian *NAGS* genes using phastCons (green) and phyloP (blue), which identified three non-coding regions of conservation located 3 kb upstream, immediately upstream, and within the first intron of *NAGS*, respectively (Figure 1). The phastCons, phyloP and our analyses of conservation within the *NAGS* gene differed due to different algorithms that were used to identify regions of conservation [39,53,54].

To validate our strategy for identification of conserved regions, the same analyses were conducted for *CPS1*, a gene in which a proximal promoter and an enhancer element located 6.3 kb upstream of rat *Cps1*, have been characterized [55,56,57]. 15 kb of *CPS1* genomic DNA sequence 5' of the translational start site was collected from human, chimpanzee, dog, mouse and rat and compared using pair-wise BLAST. Five regions of high conservation were identified including the previously reported proximal promoter located immediately upstream of the translation initiation codon and the enhancer at −7392 to −7966 relative to ATG of the human *CPS1* gene (Figure S1). In addition, three previously unknown regions, termed A, B and C, were also identified at −5, −10.5 and −12 kb relative to *CPS1* translation initiation codon (Figure S1). PhastCons and phyloP alignment of mammalian genomic DNA identified the same 5 conserved regions (Figure S1).

Highly conserved, non-coding regions of NAGS function as promoter and enhancer elements for gene transcription

Reporter assays were used to examine the functionality of each of the following: wild type *NAGS* promoter (4.10Prom), control reversed promoter (4.10PromRev), enhancer alone (4.10Enh), promoter and enhancer (4.10PromEnh), and enhancer in both orientations with the heterologous TATA-box promoter (4.23Enh and 4.23EnhRev) by measuring the expression of a luciferase reporter gene in cultured HepG2 cells (Figure 2A). Vectors pGL4.13, pGL4.23, and pGL4.10 containing firefly luciferase *luc2*,

with an SV40 promoter, a minimal TATA-promoter, or without a promoter respectively, were used as positive, baseline reference, and negative assay controls. Vector pGL4.74, containing *Renilla* luciferase *hRluc*, was co-transfected with each plasmid to control for transfection efficiency.

The human *NAGS* promoter alone (plasmid 4.10Prom), stimulated transcription of the luciferase gene while the upstream regulatory region (plasmid 4.10Enh) alone, did not (Figure 2A). When the *NAGS* promoter and upstream regulatory region were both present (4.10PromEnh plasmid), transcription increased by 50% compared to the promoter alone confirming that the upstream conserved region can function as an enhancer of transcription. When the *NAGS* enhancer was paired with a heterologous promoter containing a TATA-box, in the 4.23Enh construct, the transcription of luciferase about three times higher compared to construct with minimal TATA-box. The backbone vector 4.10 did not stimulate expression of the luciferase gene. As expected, positive control vector 4.13, containing a strong promoter, activated transcription in this cell culture system (Figure 2A). The promoter in the reverse orientation (4.10PromRev) did not activate luciferase expression indicating that the *NAGS* promoter acts in a direction dependent manner (Figure 2B). The ability of the *NAGS* enhancer (4.23EnhRev) to stimulate transcription with the heterologous promoter was orientation independent (Figure 2C). Similar results were obtained for reporter assays using mouse promoter and enhancer (Figure S2).

Transcription of NAGS initiates at multiple sites

Following discovery of the *NAGS* promoter, the transcriptional start sites (TSS) in human and mouse liver and small intestine were identified using 5' RACE (Figure 3A and B). Cloned and sequenced amplification products from 5'RACE were aligned along the 5' non-coding region of *NAGS* along with TSS identified in the Database of Transcriptional Start Sites (DBTSS) and expressed sequence tags (ESTs) from Genbank. Results suggest that *NAGS* has multiple TSS and that some may be species and tissue-specific. Combined 5'RACE, DBTSS, and Genbank results indicate that within human liver, the most frequently occurring TSS was at −42 bp upstream of the ATG codon, while in human small intestine it was at −146 bp (Figure 3A). Within mouse tissues, no dominant TSS was evident, but transcription of the *NAGS* gene initiated most often from −20 bp and −108 bp in liver and −20 bp and −95 bp in small intestine (Figure 3B). Figure 3 also shows several other rare TSS that were identified.

Transcription factors bind highly conserved motifs within the promoter and enhancer of NAGS

When promoters and enhancers from six mammalian *NAGS* genes were aligned, there were multiple regions of base pair conservation (Figure 4). Cis-eLement OVER-representation (CLOVER) software analysis was employed to identify transcription factor binding motifs in regulatory regions of human, chimpanzee, horse, cow, dog, mouse, and rat *NAGS*. Analyses of the region +9 to −996 bp (relative to the translational start codon, promoter, Table S6) and −2866 to −3620 bp (enhancer, Table S7) predicted several transcription factor binding motifs that are expressed in the liver, but no TATA-box for transcription initiation. Sp1 binding sites, within the promoter, and the HNF-1 binding motif, within the enhancer, received the highest over-representation scores, but additional motifs with lower scores were also over-represented.

Next, over-represented motifs were mapped on the CLUSTALW alignments (Figure 4A and 4B) and motifs with high conservation, having been identified in at least four out of the

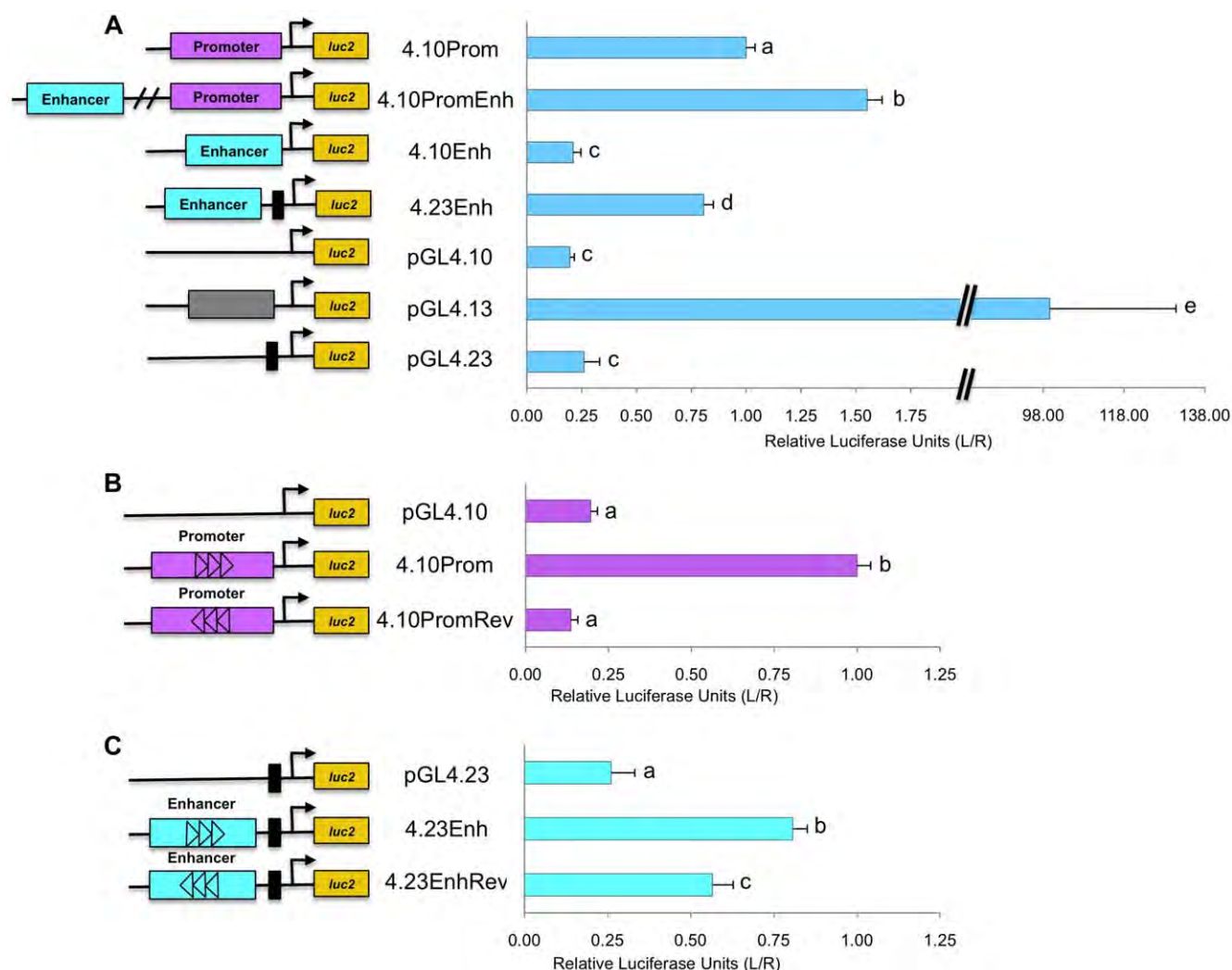


Figure 2. Highly conserved regulatory regions, upstream of the *NAGS* gene, function as promoter and enhancer elements. In liver derived cells the *NAGS* promoter (4.10Prom), promoter+enhancer (4.10PromEnh), enhancer with TATA promoter (4.23Enh), and positive control promoter vector (pGL4.13) significantly stimulate transcription while the enhancer (4.10Enh), basic vector (pGL4.10) does not stimulate transcription above baseline (A). Reverse insertion of the promoter (4.10PromRev) did not stimulate transcription compared to 4.10Prom and pGL4.10 vector (B), but reverse enhancer (4.23EnhRev) significantly stimulated transcription compared to 4.23Enh and pGL4.23 vector (C). Calculated results are an average of three independent experiments that were each carried out in triplicate, normalized to *Rluc* expression, and expressed relative to the promoter for each experiment with error reported as \pm SEM. Lowercase letters indicate statistically significant differences.
doi:10.1371/journal.pone.0029527.g002

seven mammalian species, were examined further. Throughout the promoter, five binding sites for Sp1 were highly conserved, two of which were conserved in all examined species. A binding site recognized by CREB and Activating Transcription Factor-1 (ATF-1) was conserved in four species and overlapped with the translation start codon; a C/EBP binding site was identified farther upstream in region B of the promoter (Figures 4A & 5A). Within the enhancer, a binding site for HNF-1 was conserved in all species. Overlapping binding sites for NF-Y, AP-2 and Mothers Against Decapentaplegic Homolog 3 (SMAD3) were also conserved in all species, while an additional AP-2 binding site, located 5' of the HNF-1 site, was conserved in four out of seven species (Figure 4B & 5B).

To validate computational strategy for identification of transcription factor binding sites, the enhancers of human, chimpanzee, dog, mouse, and rat *Cps1* were analyzed using CLOVER, and the experimentally identified binding motifs for

C/EBP, CREB, GR, AP-1 and HNF-3 [55,56,57] were detected along with additional unreported motifs for HNF-4, AR, C/EBP and HNF-3 (Figure S3, Table S5). The detection of experimentally confirmed binding motifs in *CPS1* has made the use of CLOVER for bioinformatic analysis of *NAGS* credible.

A DNA-protein pull-down assay was devised to test the bioinformatic prediction of specific binding sites. Two biotin-labeled DNA probes for the promoter (Figure 5A) encompassed regions A and B (Lane 1 in Figure 5C) and one probe (Figure 5B) encompassed the enhancer (Lane 1 in Figure 5D). A biotinylated probe to a region upstream of the *NAGS* gene, lacking any highly conserved motifs (Lane 3 in Figures 5C and 5D), and non-biotinylated probes to region A or B (Lane 2 in Figures 5C and 5D) were used as negative controls. The supernatant fluid from each pull-down was included as a positive control for the presence of the transcription factor (Lanes 5–8). Intensities of bands corresponding to each

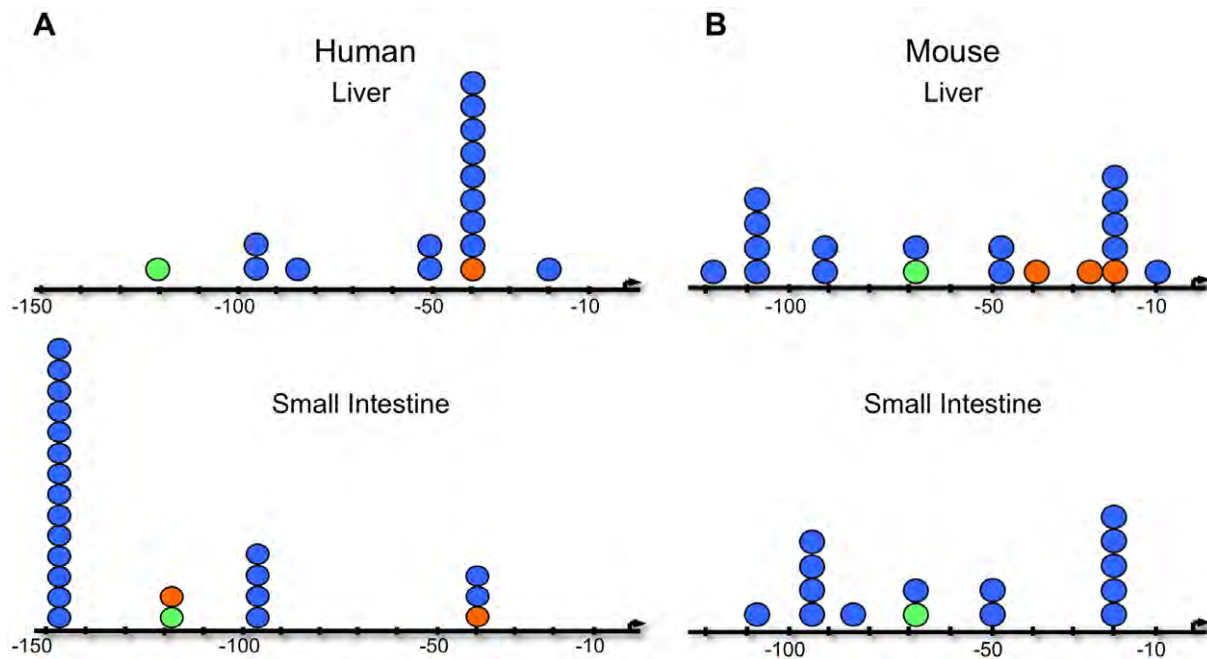


Figure 3. Transcription start sites (TSS) are species and tissue specific. TSS identified in the promoter of *NAGS* by 5'RACE analysis (blue circles), the Database of Transcriptional Start Sites (DBTSS) (green circles) and 5' termination sites of Expressed Sequence Tags (ESTs) from Genbank (orange circles) were aligned on the DNA sequence 5' of the human (A) and mouse (B) *NAGS* coding sequence. The arrow indicates the translation start site.

doi:10.1371/journal.pone.0029527.g003

transcription factor in supernatant fluids were also used as indicators of pull-down efficiency.

Factors Sp1 and CREB bound to the probe of promoter region A (Lane 1 in Figure 5C). Sp1 also bound to the probe of promoter region B (data not shown) while C/EBP did not bind to this probe (Lane 1 in Figure 5C). Within the enhancer region, transcription factors HNF-1 and NF-Y bound to the probe, however SMAD2/3 and AP2 did not (Lane 1 in Figure 5D). Binding of Sp1, CREB, C/EBP, HNF-1, NF-Y, SMAD2/3, and AP-2 was not detected in the negative controls (Lanes 2–4 in Figures 5C and 5D) while each transcription factor was detected in the positive controls of liver nuclear extract supernatants (Lanes 5–8 in Figures 5C and 5D). Each immunoblot result is representative of at least three replicate experiments.

Binding of transcription factors to the predicted motifs was also confirmed using chromatin immunoprecipitation (ChIP) followed by Real-Time PCR. Measurement compared the enrichment of target DNA regions to the negative control locus MIP-2. ChIP with Sp1 and CREB antibodies significantly enriched the *NAGS* promoter DNA compared to MIP-2 ($p < 0.005$ and $p < 0.05$, respectively; Figure 6A). ChIP with C/EBP antibody did not enrich the *NAGS* promoter DNA compared to the negative locus ($p > 0.05$; Figure 6A). The *NAGS* enhancer was enriched in chromatin immunoprecipitated with antibodies against HNF-1 and NF-Y ($p < 0.005$ and $p < 0.05$, respectively; Figure 6B), but not with antibodies against AP-2 and SMAD2/3 ($p > 0.05$; Figure 6B). Thus, Pull-down and ChIP assays confirmed that Sp1 and CREB bind along the *NAGS* promoter and HNF-1 and NF-Y bind along the enhancer.

Transcription factors and binding motifs are functionally important for transcription

Reporter assays in liver hepatoma cells with mutated transcription factor binding motifs demonstrate the functional importance of

each site. Following these sequence substitutions, transcription factor binding motifs were no longer detected by CLOVER (Table 2). Within the promoter, point mutations in the Sp1 binding sites decreased the expression of reporter gene by 75% ($p < 0.005$) and point mutations in the CREB binding site resulted in a 40% decrease ($p < 0.005$; Figure 7A). Point mutations in the HNF-1 or NF-Y binding sites, in the enhancer, decreased expression of luciferase reporter by 50% ($p < 0.005$ for both; Figure 7B).

While these results confirm that each motif is important for transcription, the functional importance of Sp1 and HNF-1 proteins is demonstrated by co-expression of the proteins with reporter assay constructs. Co-transfection of Sp1 expression plasmid with the *NAGS* promoter (4.10Prom) increases expression of luciferase more than 50% ($P < 0.005$; Figure 7A) while co-transfection of HNF-1 expression construct with the enhancer and minimal TATA promoter (4.23Enh), increases expression of the reporter gene by 25% ($p > 0.05$; Figure 7B) suggesting that endogenous Sp1 and, less so, HNF-1 do not saturate their binding motifs on the transfected reporter plasmids.

Reporter assays to compare the effect of the enhancer in liver, intestine and lung cells, included data that were normalized to the reporter expression driven by the *NAGS* promoter. While the *NAGS* enhancer (4.10PromEnh) increased expression of the reporter gene by 50% in liver derived cells (Figure 2A), expression of the luciferase gene did not increase in the intestine or lung derived cells (Figure 8) suggesting that the enhancer may determine tissue specificity of *NAGS* expression. When HNF-1 expression plasmid and 4.10PromEnh were co-transfected into intestine and lung derived cells, transcription was stimulated to levels that were not significantly different from 4.10PromEnh in liver cells ($p > 0.05$) (Figure 8). Because intestine and lung derived cells lack HNF-1 (data not shown), this demonstrated the importance of HNF-1 and *NAGS* enhancer for the tissue specificity of *NAGS* expression.

A



B.



Figure 4. Sequence alignment of *NAGS* promoters and enhancers from seven mammalian species indicate conserved motifs. DNA sequence of the promoter (A) and enhancer (B) regions were aligned using CLUSTALW alignment software. CLOVER analysis was used to identify transcription factor binding motifs. Binding sites for C/EBP (green), Sp1 (red), CREB/ATF (pink), AP-2 (purple), HNF-1 (blue), NF-Y (olive), and SMAD 3 (cyan) were highly conserved.

doi:10.1371/journal.pone.0029527.g004

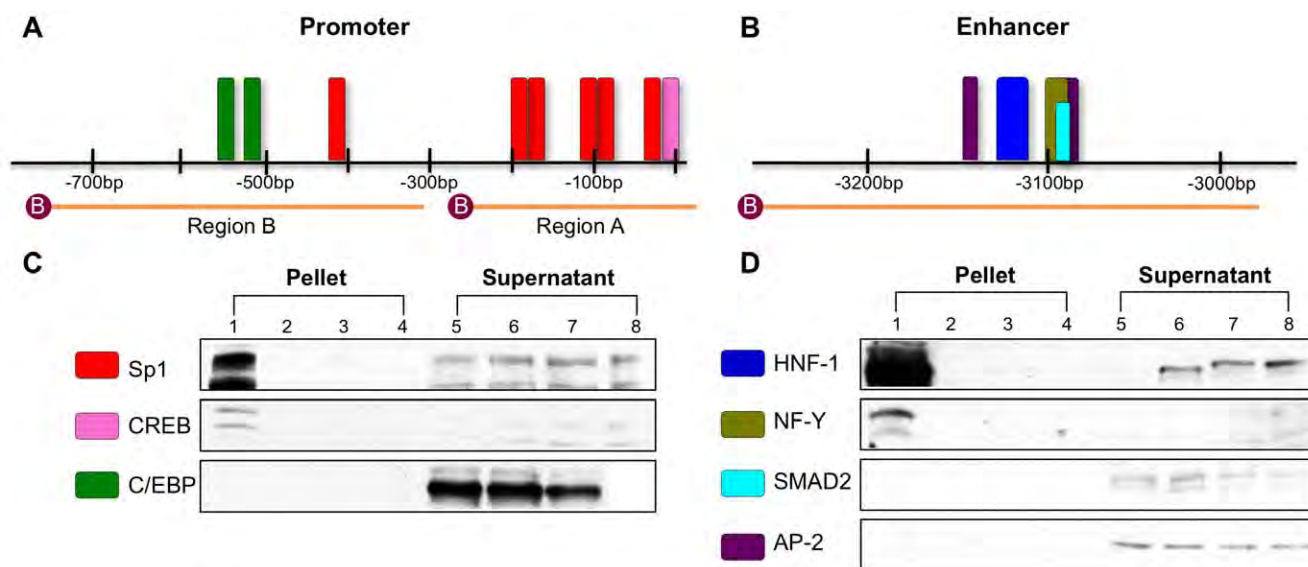


Figure 5. DNA-protein avidin-agarose pull-down assay results confirm transcription factor binding. Two probes for the promoter (A) and one probe for the enhancer (B) encompass the highly conserved transcription factor binding motifs of *NAGS*. The motif colors reflect the colors used in figures 4A and B. Assays followed by immunoblot confirmed binding of Sp1 and CREB, but not C/EBP within the promoter (C) and HNF-1 and NF-Y, but not SMAD3 or AP-2 within the enhancer regions (D). Lanes 1–4 represent precipitated proteins from mouse liver nuclear extract bound to biotinylated probes of the regions of interest (Lane 1), non-biotinylated probes of the regions of interest (Lane 2), biotinylated probes of non-specific regions (Lane 3), and no probe (Lane 4). Lanes 5–8 represent supernatant fluid from overnight incubation of biotinylated probes of the region of interest (Lane 5), non-biotinylated probes of the region of interest (Lane 6), biotinylated probes of the non-specific regions (Lane 7), or no probe (Lane 8). Immunoblots are representative of at least three replicate experiments. doi:10.1371/journal.pone.0029527.g005

Discussion

In this study we used bioinformatic analyses to predict regulatory regions based on the hypothesis that non-coding DNA sequences that are highly conserved between species are important for gene regulation. Multiple pair-wise BLAST alignments and sequence alignment from the UCSC genome browser were used to identify two conserved regions within *NAGS*, which were determined to be a promoter and an enhancer. The efficacy of this method was confirmed by successful identification of the experimentally identified promoter and -6.3 kb enhancer [25,31], along with three additional highly conserved regions, in the non-coding region upstream of *CPS1*. It should be noted that the high stringency of our BLAST analysis (80% identity and at least 100 bp of aligned sequence in four or more species) was selected to identify conserved regions that could support multiple binding sites where complexes of transcription factors may form [25,58]. This may have caused us to overlook species specific or isolated binding motifs, such as the recently identified FXR binding site [59].

The reporter assay results confirm that the two highly conserved regions within 1 kb and 3 kb upstream of the translational start site function as promoter and enhancer, respectively. The promoter activates expression of the luciferase reporter gene and we therefore infer that it will activate transcription of *NAGS* *in vivo*. Similarly, the enhancer in either orientation increases expression of luciferase by approximately 50% relative to the promoter alone, suggesting that it stimulates *NAGS* transcription as well. The relatively small but significant effect of the enhancer could be due to spacing differences between the genomic *NAGS* promoter and enhancer and their spacing in the reporter constructs. Alternatively, while HepG2 cells express transcription factors that we identified using bioinformatic tools, the *NAGS* enhancer may bind additional factors, absent in HepG2 cells, and have larger effect *in*

in vivo than in cultured cells. Another explanation for the relatively small effect of the *NAGS* enhancer is the possible presence of a proximal enhancer within the region we termed the promoter. Additional experiments are necessary to distinguish between these two possibilities.

Our analysis of the *NAGS* transcriptional start sites identified multiple TSS that may be species and tissue specific. While the function of each TSS is unknown, these results are consistent with transcription initiation by Sp1 [16,60,61], and future experiments may find that they are involved in transcriptional control for tissue specific expression, developmental-stage specific expression, quantitatively different levels of mRNA expression, or may even determine the transcript stability [62].

After we confirmed that the promoter and enhancer initiate and increase transcription, we looked for transcription factors that bind and regulate *NAGS* in these regions. By filtering for the highly over-represented and spatially conserved binding sites, relative to the translational start codon, we identified Sp1, CREB, and C/EBP in the promoter and HNF-1, AP-2, NF-Y, and SMAD-3 in the enhancer as transcription factors that could bind to the *NAGS* upstream region. This filtering method was confirmed by analysis of the -6.3 kb enhancer of *CPS1* in which binding sites for the previously published C/EBP, CREB, GR, and HNF-3 were identified.

The protein-DNA pull down assays, designed to test which transcription factors among a pool of nuclear proteins bind to amplified sequence of conserved upstream DNA, confirmed that Sp1, CREB, HNF-1 and NF-Y bind to *NAGS* promoter and enhancer, while we could not detect binding of C/EBP, AP-2 and SMAD3 (Figure 5). We initially used 60 bp probes encompassing a specific binding motif for the protein-DNA pull down assays. However, probes encompassing the entire region were better able to bind transcription factors (data not shown), suggesting that binding is facilitated by interactions with DNA sequences outside

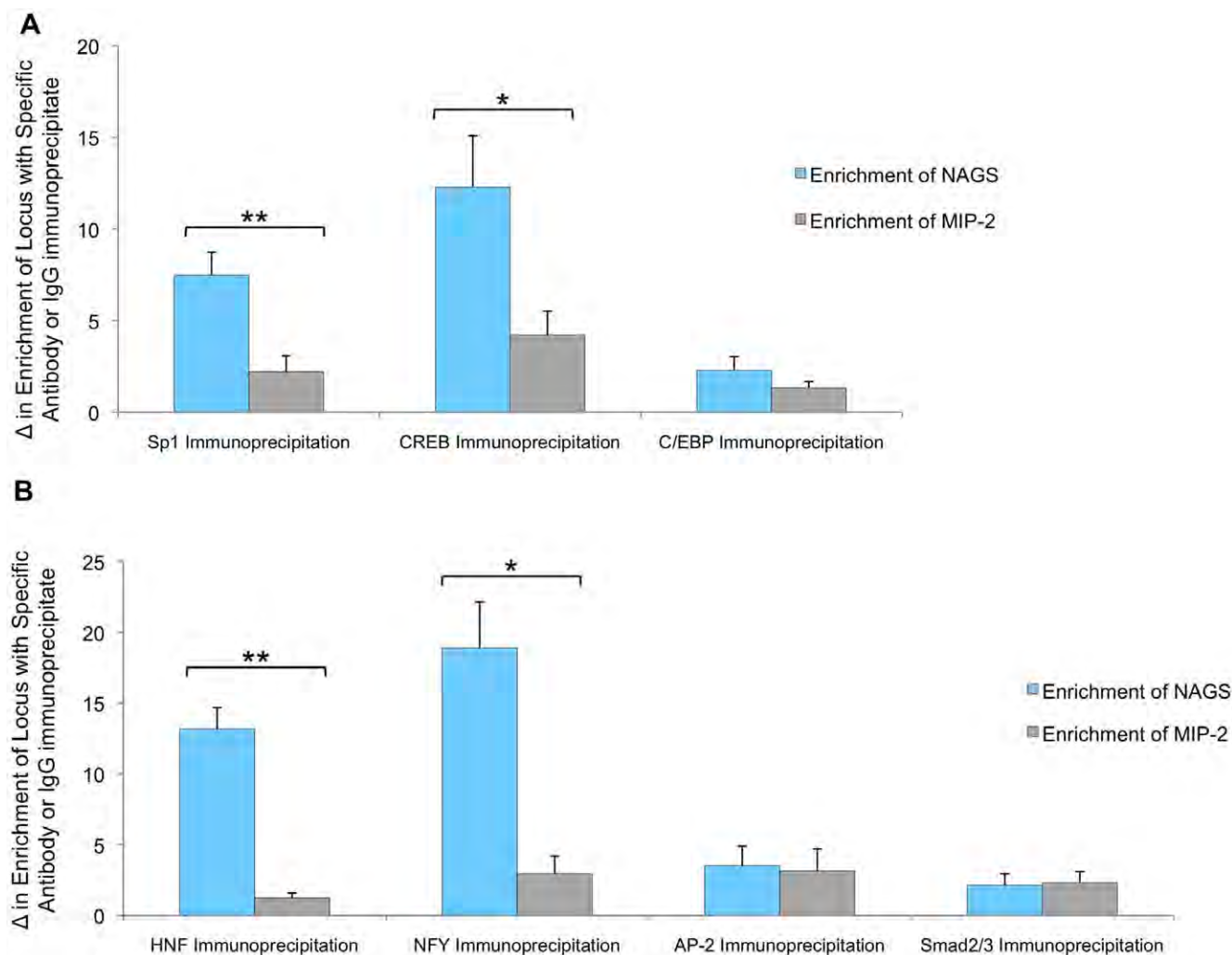


Figure 6. Chromatin Immunoprecipitation (ChIP) results confirm transcription factor binding. ChIP with transcription factor antibodies was compared to negative control IgG antibody. Real-Time PCR using promoter or enhancer specific primers was compared to primers for the negative locus MIP-2. The results confirmed that Sp1 and CREB but not C/EBP bind within the promoter (A) and HNF-1 and NF-Y but not AP-2 or SMAD2/3 bind within the enhancer region (B) of *NAGS*. Calculated error was from three replicate experiments and reported as \pm SEM. One asterisk (*) indicates $p < 0.05$ and two asterisks (**) indicate $p < 0.005$. doi:10.1371/journal.pone.0029527.g006

predicted binding sites and possibly other transcription factors and co-activators. ChIP analysis was used to confirm binding of the predicted transcription factors to the DNA regions of interest under physiological conditions. ChIP and DNA-pull down assays confirmed that Sp1 and CREB bind to the promoter and HNF-1 and NF-Y bind to the enhancer of *NAGS* (Figures 5 and 6), while reporter assays demonstrated the functional importance of each binding motif by a decrease in transcription following mutagenesis of the binding sites (Figure 7).

Furthermore, we have demonstrated that Sp1 and HNF-1 are important for stimulation of transcription of *NAGS* and that HNF-1 determines tissue specificity of *NAGS* expression. In the liver derived cell line, co-transfection of either Sp1 or HNF-1 expression plasmids with reporter constructs containing the *NAGS* promoter and enhancer led to increased expression of the reporter gene (Figure 7) suggesting that these two transcription factors regulate expression of *NAGS* in the liver. In the lung and intestine derived cell lines, expression of HNF-1 was sufficient to activate expression of reporter gene in constructs containing *NAGS*

enhancer and promoter (Figure 8). This suggests that HNF-1 binding to the *NAGS* enhancer determines tissue specificity of *NAGS* expression. Testing the effect of over-expression of CREB protein was hindered by its capacity to homo- and heterodimerize with multiple partners [63,64]. The effect of NF-Y was not tested because this transcription factor is a heterotrimer [65] and its co-expression with reporter plasmids would require stable expression of NF-Y subunit proteins by *in vitro* cell culture before reporter plasmids can be transfected and assayed for NF-Y effect on transcription.

From the data provided herein, we can speculate on the potential role these factors play in regulating *NAGS* transcription. First, in the absence of a canonical TATA-box, transcription initiated by Sp1 often results in multiple transcriptional start sites [66,67]. Sp1 is a strong activator of transcription [16,68,69,70,71] and when multiple Sp1 sites are present, as in *NAGS*, multiple Sp1 proteins can form complexes with each other and synergistically activate transcription [16,69]. Because transcription is significantly increased by co-expression with Sp1 protein and decreased

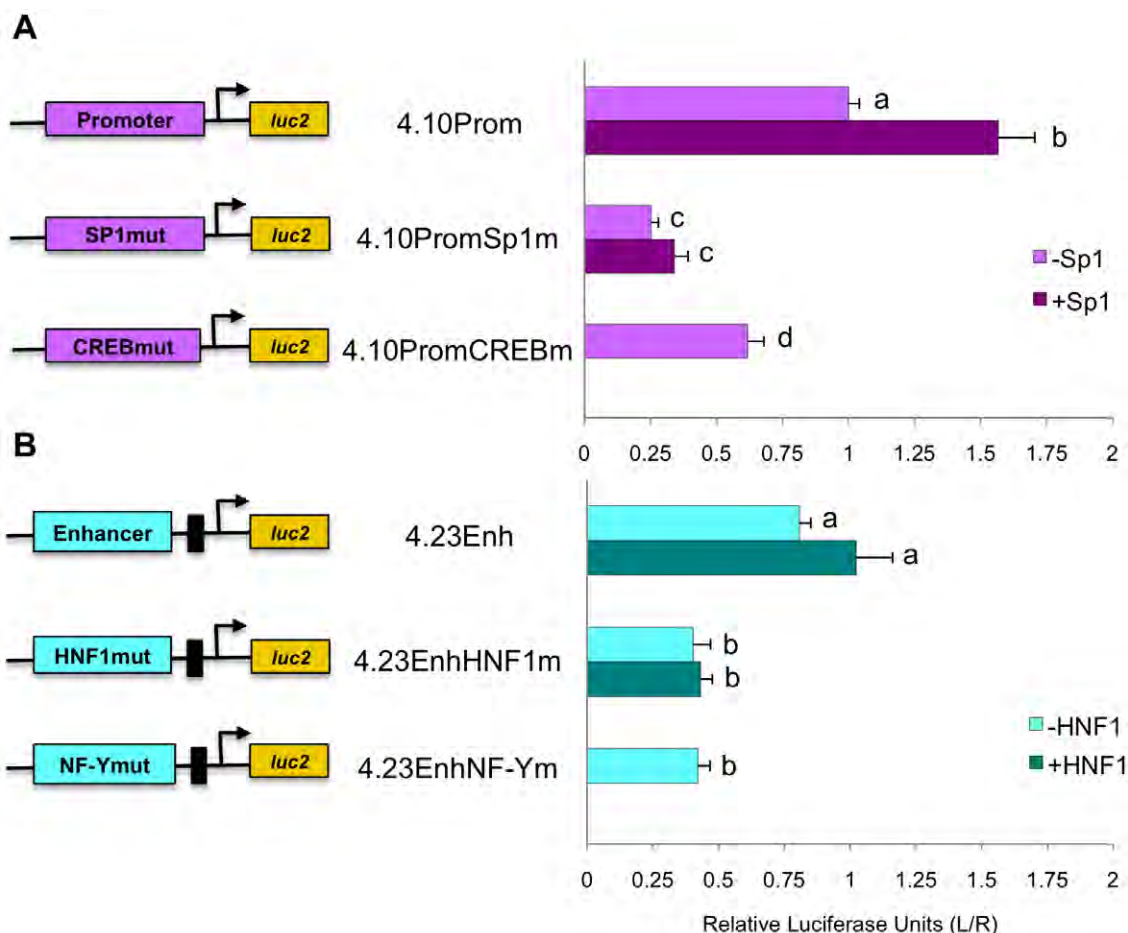


Figure 7. Transcription factors Sp1, CREB, HNF-1, and NF-Y are functionally important for stimulating expression of reporter gene transcription. Mutagenesis of the putative transcription factor binding sites significantly decreases transcription by the promoter (A) and the enhancer with TATA promoter (B) in liver derived cells when compared to non-mutated sites. Addition of Sp1 with the promoter (A) and HNF-1 with the enhancer (B) increases transcription driven by non-mutated constructs. Calculated results are an average of three independent experiments that were each carried out in triplicate, normalized to *Rluc* expression, and expressed relative to the promoter for each experiment with error reported as \pm SEM. Lowercase letters indicate statistically significant differences.
doi:10.1371/journal.pone.0029527.g007

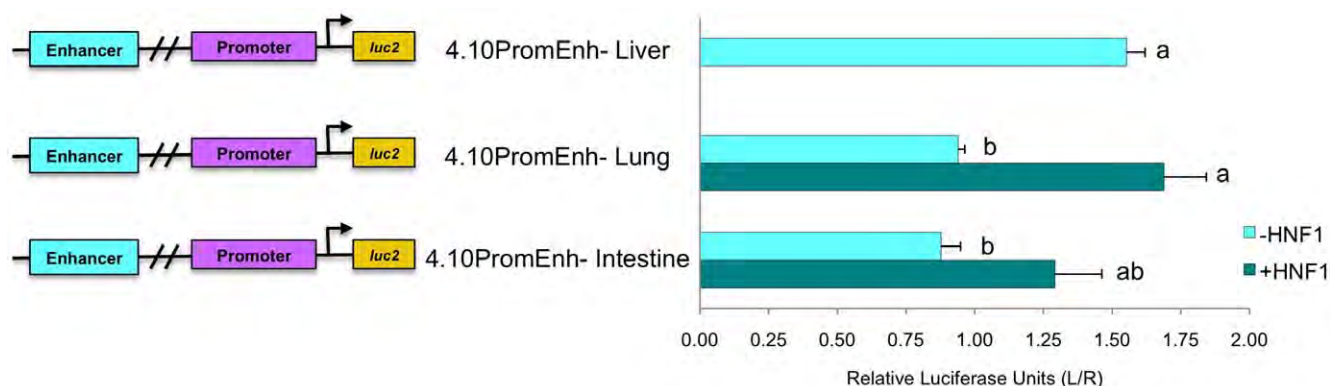


Figure 8. The NAGS enhancer shows tissue specificity. The enhancer with NAGS promoter (4.10PromEnh) increases transcription relative to the promoter in liver derived cells but not in intestine or lung derived cells (cyan bars) without the addition of HNF-1 protein (teal bars). Calculated results are an average of three independent experiments that were carried out in triplicate, normalized to *Rluc* expression, and expressed relative to the promoter for each experiment with error reported as \pm SEM. Lowercase letters indicate statistically significant differences.
doi:10.1371/journal.pone.0029527.g008

following mutation of the Sp1 binding sites, Sp1 may prove to be the activator of *NAGS* transcription, similar to its role for *ASS*, *ASL* and *ARG1* [15,25].

Second, studies have shown that glucagon and second messenger cAMP trigger a cascade that phosphorylates CREB and allows for DNA binding and activation of transcription [72,73]. In *CPS1* and *ASS*, CREB stimulates transcription upon glucagon signaling [15,31]. Decrease in transcription following CREB mutation and the close proximity of Sp1 and CREB binding sites among the TSS suggests that the transcription initiation machinery may be recruited by these factors, and future research should examine this postulate.

Our experiments and other studies [74] confirm the role of HNF-1 in *NAGS* expression. HNF-1 is essential for stimulation of *NAGS* expression by its enhancer. This factor is in part regulated by HNF-3, HNF-4, and C/EBP, each of which are known to regulate other urea cycle genes [75,76,77]. Future research will focus on the mechanism of control between these factors, HNF-1, and *NAGS*. Our study has also shown that NF-Y is an activator of *NAGS* expression, and future studies will focus on the exact mechanism of its function in this context.

The human *NAGS* gene on the forward strand of chromosome 17 partially overlaps with the peptide YY (*PYY*) gene, which is on the reverse strand. This overlap was identified with a *PYY* cDNA isolated from a brain astrocytoma cDNA library that has an 80 nucleotide long exon located between regions A and B of the *NAGS* promoter [78,79] (Figure 1). Other full-length *PYY* transcripts initiate about 500 bp upstream of the *PYY* coding region, which is located 51 kb upstream of the *NAGS* translation initiation codon. Recent analysis of human transcripts revealed that many protein coding loci are associated with at least one transcript that initiates from a distal site [80], but the significance or function of these transcripts remains to be elucidated. Partial overlap between human *NAGS* and *PYY* genes raises the interesting possibility that these two genes share *cis*-acting regulatory elements and might be co-regulated [79,81]. The mechanism of co-regulation of human *NAGS* and *PYY* is likely to be complex because of their differing tissue expression patterns [1,82,83,84] including different cell types within the intestine. *PYY* is expressed in the intestinal neuroendocrine cells [85,86] while epithelial cells in the small intestine express *NAGS* [87,88], together with *OTC* and *CPS1* [13,89]. Inspection of the transcription factor binding track of the UCSC genome browser revealed two binding sites for the CTCF transcription repressor between *NAGS* and *PYY* genes; they are located approximately 9.5 and 21 kb upstream of the *NAGS* coding region. The CTCF binding sites could act as chromatin insulators [90,91,92] and either block regulation of *PYY* by the *NAGS* enhancer or enable cell type specific regulation of each gene by the *NAGS* enhancer and promoter. Our results show that the *NAGS* promoter in the reverse orientation does not activate transcription of the reporter gene in liver derived cells (Figure 2), but this does not preclude transcription activation in other cell types, not tested in this study. It is possible that the *NAGS* promoter, enhancer, or other *NAGS* regions, regulates expression of *PYY* [84], and reporter assays in tissues and cultured cells which express *PYY* would test this hypothesis.

While regulation of *NAGS* by Sp1, CREB, HNF-1, NF-Y, and factors that regulate them, requires additional study, identification of regions that regulate human *NAGS* and *OTC* have enabled diagnosis of patients with clinical symptoms of urea cycle disorders, but lacking disease causing mutations in the coding regions of the genes [93,94]. Recently, we identified a patient with a mutation in the enhancer of *NAGS* and

confirmed the diagnosis of *NAGS* deficiency by showing that the mutation significantly decreases transcription of *NAGS* [93]. This example suggests that identification of regulatory regions within genes will lead to more and better diagnoses of urea cycle disorders and other genetic diseases and to accurate genetic counseling.

In conclusion, this study identified a promoter and a tissue specific enhancer of *NAGS* and functionally relevant transcription factor binding motifs within these regions. The results show that Sp1 and CREB bind to the *NAGS* promoter, suggesting that glucagon and cAMP signaling may regulate the expression of *NAGS*. Within the enhancer, HNF-1 may be an important factor in the coordinated regulation of this urea cycle gene transcription through its interaction with HNF-3, HNF-4 and C/EBP while the role of NF-Y is less clear considering that NF-Y may function as an activator or repressor. While additional studies will be needed to further define the roles of these factors, these results contain the first thorough analysis of *NAGS* and suggest networks of control between signaling cascades, *NAGS* and the coordinated regulation of the other urea cycle genes.

Supporting Information

Figure S1 Regions Upstream of mammalian *CPS1* genes are highly conserved. Three new highly conserved regions were identified within 15 kb 5' of the *CPS1* translational start site. Conservation algorithms phastCons (green) and phyloP (blue) from the UCSC genome browser indicate regions that are highly conserved across all mammals (A). Pair-wise blast analysis of human, chimpanzee, dog, mouse, and rat 5' non-coding region of *CPS1* were used to identify two known and three previously unknown regions of high conservation, referred to enhancer/repressor regions A, B, and C. Highly conserved regions within the *CPS1* 5' non-coding sequence include the proximal promoter, region A, the -enhancer, region B, and region C. (TIF)

Figure S2 Highly conserved regulatory regions, upstream of the mouse *Nags* gene, function as promoter and enhancer elements. Mouse promoter (m4.10Prom), promoter and enhancer (m4.10PromEnh), and enhancer with TATA promoter (m4.23Enh) stimulated transcription while enhancer lacking a promoter (m4.10Enh) did not in liver cells. Calculated results are an average of three independent experiments that were carried out in triplicate, normalized to *Rhuc* expression, and expressed relative to the promoter for each experiment with error reported as \pm SEM. (TIF)

Figure S3 Novel transcription factor binding motifs, in the enhancer region of *CPS1*, were identified using CLOVER. Several highly conserved transcription factor binding sites were present in the enhancer region. An asterisk denotes an experimentally verified transcription factor binding site. All motifs were spatially conserved between mammalian species. (TIF)

Table S1 Sequences of primers that were used to amplify human or mouse DNA by PCR for insertion of the promoter and enhancer regions into sequencing and reporter assay vectors. (DOCX)

Table S2 Primer sequences used to determine transcription start sites of *NAGS* with 5' RACE. Primers were designed according to manufacturer's instructions and used to

determine transcription start sites of human and mouse *NAGS* in liver and small intestine RNA using 5' RACE.

(DOCX)

Table S3 Primer sequences used to generate DNA probes of the specified regions of mNags. Primers were used to generate DNA probes, by PCR, of the promoter, enhancer, or non-specific specified regions of mNags.

(DOCX)

Table S4 Primer sequences used for quantitative real-time PCR analysis of chromatin immunoprecipitation samples.

(DOCX)

Table S5 Results of CLOVER analysis of the enhancer region with sequence information for human and mouse *CPS1*. Results were filtered to exclude motifs for transcription factors that are not expressed in the liver.

(DOCX)

Table S6 Results of CLOVER analysis of the promoter region with sequence information for human and mouse

References

- Caldovic L, Morizono H, Gracia Panglao M, Gallegos R, Yu X, et al. (2002) Cloning and expression of the human N-acetylglutamate synthase gene. *Biochem Biophys Res Commun* 299: 581–586.
- Caldovic L, Morizono H, Yu X, Thompson M, Shi D, et al. (2002) Identification, cloning and expression of the mouse N-acetylglutamate synthase gene. *Biochem J* 364: 825–831.
- Schimke RT (1962) Differential effects of fasting and protein-free diets on levels of urea cycle enzymes in rat liver. *J Biol Chem* 237: 1921–1924.
- Kawamoto S, Ishida H, Mori M, Tatibana M (1982) Regulation of N-acetylglutamate synthetase in mouse liver. Postprandial changes in sensitivity to activation by arginine. *Eur J Biochem* 123: 637–641.
- Tatibana M, Kawamoto S, Sonoda T, Mori M (1982) Enzyme regulation of n-acetylglutamate synthesis in mouse and rat liver. *Adv Exp Med Biol* 153: 207–216.
- Shigesada K, Tatibana M (1978) N-Acetylglutamate synthetase from rat-liver mitochondria. Partial purification and catalytic properties. *Eur J Biochem* 84: 285–291.
- Goping IS, Shore GC (1994) Interactions between repressor and anti-repressor elements in the carbamyl phosphate synthetase I promoter. *J Biol Chem* 269: 3891–3896.
- Howell BW, Lagace M, Shore GC (1989) Activity of the carbamyl phosphate synthetase I promoter in liver nuclear extracts is dependent on a cis-acting C/EBP recognition element. *Mol Cell Biol* 9: 2928–2933.
- Lagace M, Goping IS, Mueller CR, Lazzaro M, Shore GC (1992) The carbamyl phosphate synthetase promoter contains multiple binding sites for C/EBP-related proteins. *Gene* 118: 231–238.
- Schoneveld OJ, Gaemers IC, Hoogenkamp M, Lamers WH (2005) The role of proximal-enhancer elements in the glucocorticoid regulation of carbamoylphosphate synthetase gene transcription from the upstream response unit. *Biochimie* 87: 1033–1040.
- Kimura A, Nishiyori A, Murakami T, Tsukamoto T, Hata S, et al. (1993) Chicken ovalbumin upstream promoter-transcription factor (COUP-TF) represses transcription from the promoter of the gene for ornithine transcarbamylase in a manner antagonistic to hepatocyte nuclear factor-4 (HNF-4). *J Biol Chem* 268: 11125–11133.
- Kimura T, Chowdhury S, Tanaka T, Shimizu A, Iwase K, et al. (2001) CCAAT/enhancer-binding protein beta is required for activation of genes for ornithine cycle enzymes by glucocorticoids and glucagon in primary-cultured hepatocytes. *FEBS Lett* 494: 105–111.
- Murakami T, Nishiyori A, Takiguchi M, Mori M (1990) Promoter and 11-kilobase upstream enhancer elements responsible for hepatoma cell-specific expression of the rat ornithine transcarbamylase gene. *Mol Cell Biol* 10: 1180–1191.
- Nishiyori A, Tashiro H, Kimura A, Akagi K, Yamamura K, et al. (1994) Determination of tissue specificity of the enhancer by combinatorial operation of tissue-enriched transcription factors. Both HNF-4 and C/EBP beta are required for liver-specific activity of the ornithine transcarbamylase enhancer. *J Biol Chem* 269: 1323–1331.
- Guei TR, Liu MC, Yang CP, Su TS (2008) Identification of a liver-specific cAMP response element in the human argininosuccinate synthetase gene. *Biochem Biophys Res Commun* 377: 257–261.
- Anderson GM, Freytag SO (1991) Synergistic activation of a human promoter in vivo by transcription factor Sp1. *Mol Cell Biol* 11: 1935–1943.
- Boyce FM, 3rd, Pogulis RJ, Freytag SO (1989) Paradoxical regulation of human argininosuccinate synthetase cDNA minigene in opposition to endogenous gene: evidence for intragenic control sequences. *Somat Cell Mol Genet* 15: 123–129.
- Matsubasa T, Takiguchi M, Matsuda I, Mori M (1994) Rat argininosuccinate lyase promoter: the dyad-symmetric CCAAT box sequence CCAATTGG in the promoter is recognized by NF-Y. *J Biochem* 116: 1044–1055.
- Dorn A, Bollekens J, Staub A, Benoist C, Mathis D (1987) A multiplicity of CCAAT box-binding proteins. *Cell* 50: 863–872.
- Hoof van Huijsduijnen R, Li XY, Black D, Matthes H, Benoist C, et al. (1990) Co-evolution from yeast to mouse: cDNA cloning of the two NF-Y (CP-1/CBF) subunits. *EMBO J* 9: 3119–3127.
- Santoro C, Mermoud N, Andrews PC, Tjian R (1988) A family of human CCAAT-box-binding proteins active in transcription and DNA replication: cloning and expression of multiple cDNAs. *Nature* 334: 218–224.
- Takiguchi M, Mori M (1991) In vitro analysis of the rat liver-type arginase promoter. *J Biol Chem* 266: 9186–9193.
- Morris SM, Jr. (2002) Regulation of enzymes of the urea cycle and arginine metabolism. *Annu Rev Nutr* 22: 87–105.
- Morris SM, Jr., Moncman CL, Rand KD, Dizikes GJ, Cederbaum SD, et al. (1987) Regulation of mRNA levels for five urea cycle enzymes in rat liver by diet, cyclic AMP, and glucocorticoids. *Arch Biochem Biophys* 256: 343–353.
- Takiguchi M, Mori M (1995) Transcriptional regulation of genes for ornithine cycle enzymes. *Biochem J* 312(Pt 3): 649–659.
- Nebes VL, Morris SM, Jr. (1988) Regulation of messenger ribonucleic acid levels for five urea cycle enzymes in cultured rat hepatocytes. Requirements for cyclic adenosine monophosphate, glucocorticoids, and ongoing protein synthesis. *Mol Endocrinol* 2: 444–451.
- Ryall JC, Quantz MA, Shore GC (1986) Rat liver and intestinal mucosa differ in the developmental pattern and hormonal regulation of carbamoyl-phosphate synthetase I and ornithine carbamoyl transferase gene expression. *Eur J Biochem* 156: 453–458.
- Schimke RT (1963) Studies on factors affecting the levels of urea cycle enzymes in rat liver. *J Biol Chem* 238: 1012–1018.
- Hazra A, DuBois DC, Almon RR, Snyder GH, Jusko WJ (2008) Pharmacodynamic modeling of acute and chronic effects of methylprednisolone on hepatic urea cycle genes in rats. *Gene Regul Syst Bio* 2: 1–19.
- Abdullah Abu Musa DM, Kobayashi K, Yasuda I, Iijima M, Christoffels VM, et al. (1999) Involvement of a cis-acting element in the suppression of carbamoyl phosphate synthetase I gene expression in the liver of carnitine-deficient mice. *Mol Genet Metab* 68: 346–356.
- Schoneveld OJ, Hoogenkamp M, Stallen JM, Gaemers IC, Lamers WH (2007) cyclicAMP and glucocorticoid responsiveness of the rat carbamoylphosphate synthetase gene requires the interplay of upstream regulatory units. *Biochimie* 89: 574–580.
- Murakami T, Takiguchi M, Inomoto T, Yamamura K, Mori M (1989) Tissue- and developmental stage-specific expression of the rat ornithine carbamoyl-transferase gene in transgenic mice. *Dev Genet* 10: 393–401.
- Sladek FM, Zhong WM, Lai E, Darnell JE, Jr. (1990) Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. *Genes Dev* 4: 2353–2365.

NAGS. Results were filtered to exclude motifs for transcription factors that are not expressed in liver.

(DOCX)

Table S7 Results of CLOVER analysis of the enhancer region with sequence information for human and mouse

NAGS. Results were filtered to exclude motifs for transcription factors that are not expressed in the liver.

(DOCX)

Acknowledgments

We would like to thank Dr. Marshall Summar for providing HepG2 cells and Dr. Mary Rose for providing A549 cells.

Author Contributions

Analyzed the data: SKH. Contributed reagents/materials/analysis tools: LM-R. Wrote the paper: SKH. Designed the experiments: SKH. Critically reviewed the manuscript: MT. Conceived the study and reviewed the manuscript: LC. Performed reporter assays, 5'-RACE, DNA pull-down assays, ChIP, bioinformatic analysis and wrote the paper: SKH. Carried out and analyzed bioinformatic analysis of the NAGS upstream regulatory region: GYL MP SS LM-R.

34. Boyce FM, Anderson GM, Rusk CD, Freytag SO (1986) Human argininosuccinate synthetase minigenes are subject to arginine-mediated repression but not to trans induction. *Mol Cell Biol* 6: 1244–1252.
35. Boyce FM, 3rd, Freytag SO (1989) Regulation of human argininosuccinate synthetase gene: induction by positive-acting nuclear mechanism in canavanine-resistant cell variants. *Somat Cell Mol Genet* 15: 113–121.
36. Jackson MJ, Allen SJ, Beaudet AL, O'Brien WE (1988) Metabolite regulation of argininosuccinate synthetase in cultured human cells. *J Biol Chem* 263: 16388–16394.
37. Sunyakumthorn P, Boonsaen T, Boonsaeng V, Wallace JC, Jitrapakdee S (2005) Involvement of specific proteins (Sp1/Sp3) and nuclear factor Y in basal transcription of the distal promoter of the rat pyruvate carboxylase gene in beta-cells. *Biochem Biophys Res Commun* 329: 188–196.
38. Snodgrass PJ (1991) Dexamethasone and glucagon cause synergistic increases of urea cycle enzyme activities in livers of normal but not adrenalectomized rats. *Enzyme* 45: 30–38.
39. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.
40. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, et al. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* 32: 1372–1381.
41. Marino-Ramirez L, Spouge JL, Kanga GC, Landsman D (2004) Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res* 32: 949–958.
42. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, et al. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 31: 3497–3500.
43. Kim HJ, Ko MS, Kim HK, Cho WJ, Lee SH, et al. Transcription factor Sp1 regulates basal transcription of the human DRG2 gene. *Biochim Biophys Acta* 1809: 184–190.
44. Zhang W, Tian Z, Sha S, Cheng LY, Philipsen S, et al. Functional and sequence analysis of human neuroglobin gene promoter region. *Biochim Biophys Acta* 1809: 236–244.
45. Convertini P, Infantino V, Bisaccia F, Palmieri F, Iacobazzi V. Role of FOXA and Sp1 in mitochondrial acylcarnitine carrier gene expression in different cell lines. *Biochem Biophys Res Commun* 404: 376–381.
46. Michels AJ, Hagen TM (2009) Hepatocyte nuclear factor 1 is essential for transcription of sodium-dependent vitamin C transporter protein 1. *Am J Physiol Cell Physiol* 297: C1220–1227.
47. Wang Z, Burke PA. Hepatocyte nuclear factor-4alpha interacts with other hepatocyte nuclear factors in regulating transthyretin gene expression. *FEBS J* 277: 4066–4075.
48. Tue NT, Yoshioka Y, Yamaguchi M. NF-Y transcriptionally regulates the *Drosophila* p53 gene. *Gene* 473: 1–7.
49. Pallai R, Simpkins H, Chen J, Parekh HK. The CCAAT box binding transcription factor, nuclear factor-Y (NF-Y) regulates transcription of human aldo-keto reductase 1C1 (AKR1C1) gene. *Gene* 459: 11–23.
50. Xiang H, Wang J, Boxer LM (2006) Role of the cyclic AMP response element in the bcl-2 promoter in the regulation of endogenous Bcl-2 expression and apoptosis in murine B cells. *Mol Cell Biol* 26: 8599–8606.
51. Callens N, Baert JL, Monte D, Sunesen M, Van Lint C, et al. (2003) Transcriptional regulation of the murine *brca2* gene by CREB/ATF transcription factors. *Biochem Biophys Res Commun* 312: 702–707.
52. Deng WG, Zhu Y, Montero A, Wu KK (2003) Quantitative analysis of binding of transcription factor complex to biotinylated DNA probe by a streptavidin-agarose pulldown assay. *Anal Biochem* 323: 12–18.
53. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 20: 110–121.
54. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15: 1034–1050.
55. Christoffels VM, van den Hoff MJ, Lamers MC, van Roon MA, de Boer PA, et al. (1996) The upstream regulatory region of the carbamoyl-phosphate synthetase I gene controls its tissue-specific, developmental, and hormonal regulation in vivo. *J Biol Chem* 271: 31243–31250.
56. Christoffels VM, Grange T, Kaestner KH, Cole TJ, Darlington GJ, et al. (1998) Glucocorticoid receptor, C/EBP, HNF3, and protein kinase A coordinately activate the glucocorticoid response unit of the carbamoylphosphate synthetase I gene. *Mol Cell Biol* 18: 6305–6315.
57. Christoffels VM, van den Hoff MJ, Moorman AF, Lamers WH (1995) The far-upstream enhancer of the carbamoyl-phosphate synthetase I gene is responsible for the tissue specificity and hormone inducibility of its expression. *J Biol Chem* 270: 24932–24940.
58. Klein H, Vingron M (2007) Using transcription factor binding site co-occurrence to predict regulatory regions. *Genome Inform* 18: 109–118.
59. Renga B, Mencarelli A, Cipriani S, D'Amore C, Zampella A, et al. The nuclear receptor FXR regulates hepatic transport and metabolism of glutamine and glutamate. *Biochim Biophys Acta* 1812: 1522–1531.
60. Emami KH, Burke TW, Smale ST (1998) Sp1 activation of a TATA-less promoter requires a species-specific interaction involving transcription factor IID. *Nucleic Acids Res* 26: 839–846.
61. Muckenfuss H, Kaiser JK, Krebil E, Battenberg M, Schwer C, et al. (2007) Sp1 and Sp3 regulate basal transcription of the human APOBEC3G gene. *Nucleic Acids Res* 35: 3784–3796.
62. Schibler U, Sierra F (1987) Alternative promoters in developmental gene expression. *Annu Rev Genet* 21: 237–257.
63. Hai T, Hartman MG (2001) The molecular biology and nomenclature of the activating transcription factor/cAMP responsive element binding family of transcription factors: activating transcription factor proteins and homeostasis. *Gene* 273: 1–11.
64. De Cesare D, Sassone-Corsi P (2000) Transcriptional regulation by cyclic AMP-responsive factors. *Prog Nucleic Acid Res Mol Biol* 64: 343–369.
65. Matuoka K, Yu Chen K (1999) Nuclear factor Y (NF-Y) and cellular senescence. *Exp Cell Res* 253: 365–371.
66. Juang HH, Costello LC, Franklin RB (1995) Androgen modulation of multiple transcription start sites of the mitochondrial aspartate aminotransferase gene in rat prostate. *J Biol Chem* 270: 12629–12634.
67. Pave-Preux M, Aggerbeck M, Veyssier C, Bousquet-Lemerrier B, Hanoune J, et al. (1990) Hormonal discrimination among transcription start sites of aspartate aminotransferase. *J Biol Chem* 265: 4444–4448.
68. Kadonaga JT, Courey AJ, Ladika J, Tjian R (1988) Distinct regions of Sp1 modulate DNA binding and transcriptional activation. *Science* 242: 1566–1570.
69. Li L, He S, Sun JM, Davie JR (2004) Gene regulation by Sp1 and Sp3. *Biochem Cell Biol* 82: 460–471.
70. Solomon SS, Majumdar G, Martinez-Hernandez A, Raghov R (2008) A critical role of Sp1 transcription factor in regulating gene expression in response to insulin and other hormones. *Life Sci* 83: 305–312.
71. Wierstra I (2008) Sp1: emerging roles—beyond constitutive activation of TATA-less housekeeping genes. *Biochem Biophys Res Commun* 372: 1–13.
72. Montminy M, Koo SH, Zhang X (2004) The CREB family: key regulators of hepatic metabolism. *Ann Endocrinol (Paris)* 65: 73–75.
73. Mayr B, Montminy M (2001) Transcriptional regulation by the phosphorylation-dependent factor CREB. *Nat Rev Mol Cell Biol* 2: 599–609.
74. Heibel SK, Ah Mew N, Caldovic L, Daikhin Y, Yudkoff M, et al. N-carbamylglutamate enhancement of ureagenesis leads to discovery of a novel deleterious mutation in a newly defined enhancer of the NAGS gene and to effective therapy. *Hum Mutat* 32: 1153–1160.
75. Kuo CJ, Conley PB, Chen L, Sladek FM, Darnell JE, Jr., et al. (1992) A transcriptional hierarchy involved in mammalian cell-type specification. *Nature* 355: 457–461.
76. Sladek FM (1993) Orphan receptor HNF-4 and liver-specific gene expression. *Receptor* 3: 223–232.
77. Kistaki E, Talianidis I (1997) Modulation of hepatic gene expression by hepatocyte nuclear factor 1. *Science* 277: 109–112.
78. Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, et al. (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci U S A* 99: 16899–16903.
79. Lomenick JP, Melguizo MS, Mitchell SL, Summar ML, Anderson JW (2009) Effects of meals high in carbohydrate, protein, and fat on ghrelin and peptide YY secretion in prepubertal children. *J Clin Endocrinol Metab* 94: 4463–4471.
80. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799–816.
81. Mitchell S, Murdock D, Summar M (2008) Plasma peptide tyrosine tyrosine (PYY) levels are increased in urea cycle disorder patients. *Mol Gen Metab* 93: 258 (abstract).
82. Neill MA, Aschner J, Barr F, Summar ML (2009) Quantitative RT-PCR comparison of the urea and nitric oxide cycle gene transcripts in adult human tissues. *Mol Genet Metab* 97: 121–127.
83. Myrden-Axcrone U, Ekblad E, Sundler F (1997) Developmental expression of NPY, PYY and PP in the rat pancreas and their coexistence with islet hormones. *Regul Pept* 68: 165–175.
84. Ekblad E, Sundler F (2002) Distribution of pancreatic polypeptide and peptide YY. *Peptides* 23: 251–261.
85. Lundberg JM, Tatemoto K, Terenius L, Hellstrom PM, Mutt V, et al. (1982) Localization of peptide YY (PYY) in gastrointestinal endocrine cells and effects on intestinal blood flow and motility. *Proc Natl Acad Sci U S A* 79: 4471–4475.
86. Lukinius AI, Ericsson JL, Lundqvist MK, Wilander EM (1986) Ultrastructural localization of serotonin and polypeptide YY (PYY) in endocrine cells of the human rectum. *J Histochem Cytochem* 34: 719–726.
87. Geng M, Li T, Kong X, Song X, Chu W, et al. Reduced expression of intestinal N-acetylglutamate synthase in suckling piglets: a novel molecular mechanism for arginine as a nutritionally essential amino acid for neonates. *Amino Acids* 40: 1513–1522.
88. Uchiyama C, Mori M, Tatibana M (1981) Subcellular localization and properties of N-acetylglutamate synthase in rat small intestinal mucosa. *J Biochem* 89: 1777–1786.
89. Dubois N, Cavard C, Chasse JF, Kamoun P, Briand P (1988) Compared expression levels of ornithine transcarbamylase and carbamylphosphate synthetase in liver and small intestine of normal and mutant mice. *Biochim Biophys Acta* 950: 321–328.
90. Ishihara K, Oshimura M, Nakao M (2006) CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Mol Cell* 23: 733–742.
91. Renda M, Baglivo I, Burgess-Beusse B, Esposito S, Fattorusso R, et al. (2007) Critical DNA binding interactions of the insulator protein CTCF: a small

- number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *J Biol Chem* 282: 33336–33345.
92. Majumder P, Gomez JA, Chadwick BP, Boss JM (2008) The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. *J Exp Med* 205: 785–798.
 93. Heibel SK, Ah Mew N, Caldovic L, Daikhin Y, Yudkoff M, et al. (2011) N-carbamylglutamate enhancement of ureagenesis leads to discovery of a novel deleterious mutation in a newly defined enhancer of the NAGS gene and to effective therapy. *Hum Mutat* 32: 1153–1160.
 94. Luksan O, Jirsa M, Eberova J, Minks J, Treslova H, et al. (2010) Disruption of OTC promoter-enhancer interaction in a patient with symptoms of ornithine carbamoyltransferase deficiency. *Hum Mutat* 31: E1294–1303.

RESEARCH ARTICLE

Open Access

In silico identification and characterization of the ion transport specificity for P-type ATPases in the *Mycobacterium tuberculosis* complex

Lorena Novoa-Aponte¹, Andrés León-Torres¹, Miyer Patiño-Ruiz¹, Jenifer Cuesta-Bernal¹, Luz-Mary Salazar¹, David Landsman², Leonardo Mariño-Ramírez^{2,3*} and Carlos-Yesid Soto¹

Abstract

Background: P-type ATPases hydrolyze ATP and release energy that is used in the transport of ions against electrochemical gradients across plasma membranes, making these proteins essential for cell viability. Currently, the distribution and function of these ion transporters in mycobacteria are poorly understood.

Results: In this study, probabilistic profiles were constructed based on hidden Markov models to identify and classify P-type ATPases in the *Mycobacterium tuberculosis* complex (MTBC) according to the type of ion transported across the plasma membrane. Topology, hydrophobicity profiles and conserved motifs were analyzed to correlate amino acid sequences of P-type ATPases and ion transport specificity. Twelve candidate P-type ATPases annotated in the *M. tuberculosis* H37Rv proteome were identified in all members of the MTBC, and probabilistic profiles classified them into one of the following three groups: heavy metal cation transporters, alkaline and alkaline earth metal cation transporters, and the beta subunit of a prokaryotic potassium pump. Interestingly, counterparts of the non-catalytic beta subunits of Hydrogen/Potassium and Sodium/Potassium P-type ATPases were not found.

Conclusions: The high content of heavy metal transporters found in the MTBC suggests that they could play an important role in the ability of *M. tuberculosis* to survive inside macrophages, where tubercle bacilli face high levels of toxic metals. Finally, the results obtained in this work provide a starting point for experimental studies that may elucidate the ion specificity of the MTBC P-type ATPases and their role in mycobacterial infections.

Keywords: Tuberculosis, *Mycobacterium tuberculosis* complex, P-type ATPases, Ion transport, Conserved motifs

Background

Tuberculosis (TB) is one of the most important challenges in public health maintenance throughout the world. According to the World Health Organization (WHO), 8.5-9.2 million new TB cases were estimated to have occurred in 2010 [1], and 1.2-1.5 million deaths were caused by species of the *Mycobacterium tuberculosis* complex (MTBC) that includes *M. tuberculosis*, *M. bovis*, *M. bovis* BCG (vaccine strain), *M. africanum*, *M. microti*, *M. canettii*, and *M. pinnipedii*, which produces TB in humans and some animal hosts [2,3]. Part of the infected population will develop active TB,

whereas the majority of cases (approximately 90%) progress to a non-infectious disease or latent TB, where mycobacteria survive in a dormant state inside immune cells [4]. Individuals with latent TB may be asymptomatic during prolonged periods of time; however, TB can reactivate when the host immune response diminishes due to malnutrition, steroid use, and HIV co-infection [5].

The emergence of multidrug and extensively drug-resistant tuberculosis strains (MDR-TB and XDR-TB) and the lack of drugs against latent TB have become serious problems for TB control. Therefore, the identification of new therapeutic targets useful in the development of novel drugs and vaccines against latent TB is essential. New anti-TB drugs, such as diarylquinolines (TMC207) and benzothiazines (BTZ043) target

* Correspondence: marino@ncbi.nlm.nih.gov

²Computational Biology Branch, NCBI, NLM, NIH, Bethesda, USA

³PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia
Full list of author information is available at the end of the article

essential membrane proteins that affect mycobacterial viability [6]. Thus, antimicrobials designed against proteins of plasma membrane are ideal because they avoid problems related to membrane permeability.

Ion transport in bacteria is carried out by enzymatic systems that belong to either the P-type ATPase, ATP binding cassettes (ABC transporters) and metallic ion/H⁺-antiporter systems [7]. In general, ATPases help maintain the ion gradients responsible for cell volume control and transport of nutrients across the cell membrane [8-11]. ATPases hydrolyze ATP, releasing energy that is used in the transport of ions against electrochemical gradients in plasma membranes. The enzymatic mechanisms of P-type ATPases were initially described in eukaryotic cells [12-14]. These enzymes have five different functional and structural domains: three of these domains are cytoplasmic (A, actuator; N, nucleotide binding and P, phosphorylation), and the other two are embedded in the membrane (T, transport and S, class specific support domain) [15]. P-type ATPases have the following two conformational states: E1, which binds ion substrates on one side of the cell membrane, inducing their autophosphorylation and generating a new conformational state; and E2, which has a lower affinity for substrates and therefore releases them to the other side of the cell membrane, promoting ion transport and finally recovery of the E1 state [15,16].

Despite P-type ATPases share the same catalytic mechanism based on conformational changes in their five structural domains, their regulation and substrate affinities are different [15,16]. P-type ATPases are phylogenetically classified into five subfamilies (P_I-P_V), and within these subfamilies are 10 different subtypes that are categorized based on the transported substrate [17].

In this study, probabilistic profiles were constructed to compare and classify all P-type ATPases of the MTBC based on their structural features and ion transport. Twelve possible P-type ATPases were detected in the proteome of *M. tuberculosis* H37Rv and from other members of the MTBC. The high number of heavy metal transporters discovered in the MTBC suggests an important role for P-type ATPases in *M. tuberculosis* survival within macrophages.

Methods

Construction of hidden Markov models (HMM)

To obtain a representative group of each phylogenetic subfamily, a set of 128 well-characterized P-type ATPases with evidence of existence at the protein level were retrieved from *Uniprot* (Swiss-Prot section) [18]. Each group of sequences was aligned using the *Praline* tool (<http://www.ibi.vu.nl/programs/pralinewww/>) with the BLOSUM62 matrix and *Phobius* transmembrane structure predictor, which was developed to improve the

multiple alignments of membrane protein sequences [19]. The *HMM* package of programs [20], available in the *Mobyle Pasteur* portal (<http://mobyle.pasteur.fr/cgi-bin/portal.py#welcome>), was used to find these types of pumps in the MTBC proteomes. The *Hmmbuild* tool with default settings and the multiple sequence alignments was used for HMM building. The default parameters of the *Hmmer* tools use an ad hoc position-based sequence weighting algorithm that makes the models appropriate for the identification of distant members of the P-type ATPases family. Consensus sequences were generated with the *Hmmemit* tool.

Search and classification of MTBC P-type ATPases

To date, the following 10 MTBC genomes have been completely sequenced and assembled (NCBI): *M. africanum* GM041182 (NC_015758), *M. bovis* AF2122/97 (NC_002945), *M. bovis* BCG str. Pasteur 1173P2 (NC_008769), *M. bovis* BCG str. Tokyo 172 (NC_012207), *M. canettii* CIPT 140010059 (NC_015848) and five *M. tuberculosis* strains, H37Rv (NC_000962), H37Ra (NC_009525), F11 (NC_009565), CDC1551 (NC_002755) and KZN1435 (NC_012943); these proteomes were obtained from *Uniprot* (<http://www.uniprot.org/>). Because strains of *M. microti* and *M. pinnipedii* had not been sequenced, they were not included in this study. HMM and *Hmmsearch* tool were used to find P-type ATPases in the MTBC proteomes.

Topology prediction

The topology derived from predictions of transmembrane segments (TMS) was made with the following six programs: *TopPred* (<http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::toppred>), *DAS* (www.sbc.su.se/~miklos/DAS/), *TMpred* (www.ch.embnet.org/software/TMPRED_form.html), *TMHMM 2.0* (www.cbs.dtu.dk/services/TMHMM), *HMMTOP* (www.enzim.hu/hmmtop/) and *Phobius* (phobius.sbc.su.se/). All programs were used with default settings, except *TopPred*, which allows the user to provide information about the type of organism. *TMDet* and *PPM* servers were used in cases where the predictions based on amino acid sequences did not yield reliable results. These tools must be fed with PDB files to identify TMS based on tertiary protein structure and to generate 3D models with TMS located into hypothetical lipid bilayer planes. Tertiary structure models of CtpH and CtpI were made with the *I-TASSER* tool [21] based on the *threading* strategy and were validated with the *Whatif* package of programs (<http://swift.cmbi.ru.nl/servers/html/index.html>).

Hydrophobicity profile construction

The amino acid sequence of the *M. tuberculosis* H37Rv P-type ATPases and consensus sequences generated by *Hmmemit* tool were analyzed by *TMHMM 2.0*.

Hydrophobicity profiles of consensus sequences for previously characterized P-type ATPases were used as comparison patterns.

Conserved motif analysis

Amino acids sequences of the identified proteins were manually analyzed to determine the nine conserved motifs typical of the P-type ATPase family, as previously described by Thever *et al.* [22] and others [16,23].

Results and discussion

An unusually high number of cation transporter P-type ATPases are present in the *Mycobacterium tuberculosis* complex

Multiple alignments of 128 reported P-type ATPase protein sequences from a representative group of eukaryotic and prokaryotic cells (obtained from a curated database and confirmed at the protein level) allowed the identification of highly conserved regions within the family and the classification of these members according to ion transport. These alignments were used as the starting point for the construction of HMM profiles that represented groups of ion transporters, which were also used to generate a consensus sequence for each group. The designed HMM were then used to identify P-type ATPases in proteomes of the different MTBC species, whose genome sequences have been reported. The proposed classification for the studied sequences was based on the obtained scores using the HMM and the *Hmmsearch* tool.

Because the P-type ATPases are transmembrane proteins, typical alignments using BLOSUM and PAM matrices were not adequate. In this study, the *Praline* tool was used because it considers the differences in evolutionary tendencies of transmembranal and non-transmembranal regions. *Praline* applies an adequate matrix for each protein region based on the previous prediction of TMS (made with the *Phobius* algorithm in this case). As P-type ATPases contain ion-binding motifs within a transmembrane domain, the strategy that we employed produced more reliable alignments. Twelve hypothetical proteins with a high probability of being P-type ATPase transporters of metallic cations were identified in each of the MTBC proteomes. All proteins identified in the MTBC share at least 98% identity with their orthologs in the scanned proteomes; however, *M. canettii* and *M. africanum* displayed slight differences in non-conserved regions. The pumps identified in the *M. tuberculosis* H37Rv proteome, CtpA, CtpB, CtpC, CtpD, CtpE, CtpF, CtpG, CtpH, CtpI, CtpJ, CtpV and KdpB, were in agreement with the automatic annotation of probable cation transporter P-type ATPases in the *M. tuberculosis* H37Rv genome [24] and were taken as references to facilitate further analysis of results.

A broad diversity of cations are potentially transported by *Mycobacterium tuberculosis* P-type ATPases

Figure 1 shows the *Hmmsearch* scores obtained when the 16 HMM were used to find P-type ATPases in the *M. tuberculosis* H37Rv proteome. Those scores show the similarity between the identified sequences in the H37Rv proteome and the grouped sequences used for each HMM construction. The patterns thus obtained allowed classification of the 12 P-type ATPases into the following three groups: transporters of heavy metal (HM) cations, transporters of alkaline and alkaline earth metal (AEM) cations, and a group composed only of the β subunit of the prokaryotic K^+ transporter, KdpB. The highest similarity with the conventional P-type ATPases used in this study corresponded to KdpB, CtpF and CtpV, whereas CtpE, CtpH and CtpI had the lowest scores.

The HM group is composed of CtpA, CtpB, CtpC, CtpD, CtpG, CtpJ and CtpV, which may transport Cu^{2+} , Cu^+ , Co^{2+} , Ag^+ , Hg^{2+} , Cd^{2+} , Pb^{2+} and Zn^{2+} . Most of the proteins analyzed in this work correspond to the HM group (60%) suggesting that the active transport of heavy metal cations is relevant for the tubercle bacilli persistence, as it has been hypothesized for other prokaryotes and some unicellular eukaryotes [23]. Interestingly, evidence of toxic concentrations of intracellular heavy metals has been recently described in macrophages infected with *M. tuberculosis* [25].

Alternatively, CtpE, CtpF, CtpH and CtpI belong to the AEM group and may be involved in Na^+ , K^+ , Ca^{2+} , H^+ and Mg^{2+} transport. It is possible that some of these proteins correspond to Na^+/K^+ or H^+/K^+ ATPase pumps, but they only appeared to contain the α subunits in their tridimensional structure. Their non-catalytic- β -subunit counterpart, which has been correlated with regulatory processes and assembly of P-type ATPases into the cell membrane of eukaryotic cells [10], was not found in any of the MTBC proteomes.

KdpB is very different protein from the other MTBC ATPases. It shares 63% identity with the KdpB of *E. coli* (P03960), which corresponds to the β subunit of a K^+ transporting multimeric ATPase [22,26]. Despite the observation that KdpB has the characteristic DKTGTLT phosphorylation motif of this type of pump, it does not have a typical ion binding motif [26]. In conclusion, the proposed classification for P-type ATPases of the MTBC provides an initial approximation of their functional characterization. It is noteworthy that the constructed HMM in this work became a useful tool for the identification of P-type ATPases in other biological systems.

Three different topologies can be adopted by P-type ATPases in *Mycobacterium tuberculosis*

The six different algorithms used in the hydrophobicity analysis showed that all of the *M. tuberculosis* H37Rv

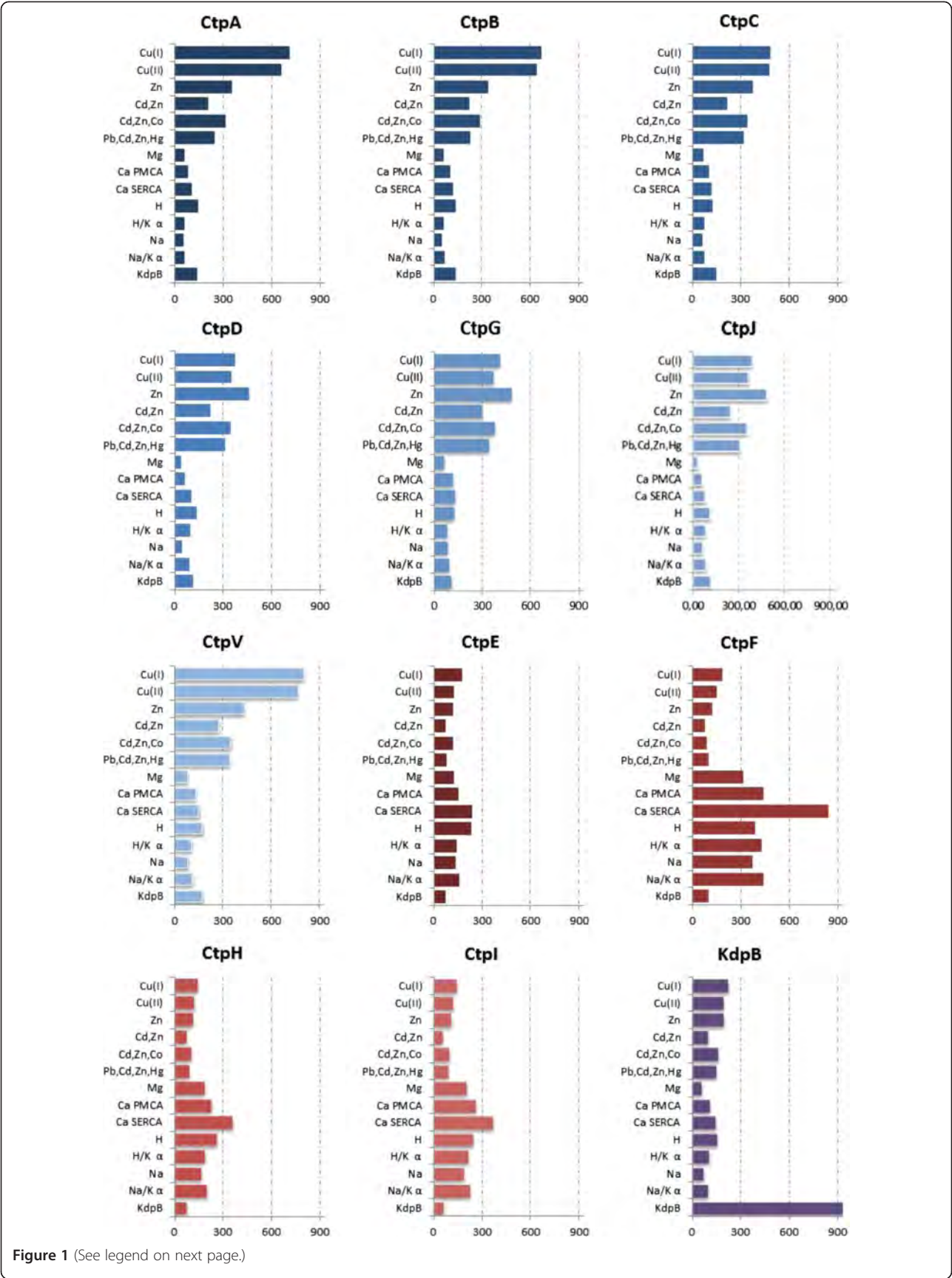


Figure 1 (See legend on next page.)

(See figure on previous page.)

Figure 1 Classification of P-type ATPases of *M. tuberculosis* H37Rv according to ion specificity. Amino acid sequences of *M. tuberculosis* H37Rv P-type ATPases were compared with the HMM profiles built from characterized P-type ATPases reported by *Uniprot*. Grouping is based on the scores obtained from *Hmmsearch*. HM group (blue bars) includes the P-type ATPase transporters of heavy metals (CtpA, CtpB, CtpC, CtpD, CtpG, CtpJ, CtpV). The EAM group (red bars) includes transporters of alkaline and alkaline earth metals (CtpE, CtpF, CtpH, CtpI). The KdpB group (purple bars) includes the β subunit of K^+ P-type ATPase. The scores obtained from the *Hmmsearch* tool are shown in the bottom of each plot.

P-type ATPases have an α -helix type TMS containing at least 17 amino acid residues, and the agreement of the topology results obtained using these prediction tools increases the confidence in the predicted TMS. Consensus transmembrane regions for each P-type ATPase were obtained if at least four of the algorithms gave similar results. Because three of the tools (*TMHMM 2.0*, *HMMTOP* and *Phobius*) incorporate size and composition restrictions in TMS, the results obtained using these algorithms are significant.

The strategy used for topology analysis showed the following three different types of topology within the *M. tuberculosis* H37Rv P-type ATPases: type I, which corresponds to HM pumps with eight TMS (A, B and from 1 to 6), type II, which corresponds to AEM pumps with 10 TMS (from 1 to 10), and type III, which corresponds to KdpB with 7 TMS (Figure 2). It was observed that all P-type ATPases have two cytoplasmic loops (small and large) that include the phosphorylation and ATP binding sites. The small cytoplasmic loop is located between TMS2 and TMS3, whereas the largest cytoplasmic loop is situated between TMS4 and TMS5 (Figure 2) [22]. Additionally, as was expected, the N- and C-termini of these proteins are located in the cytosolic side, except in the case of KdpB, in which the N-terminus is located intracellularly and the C-terminus is outside the cytoplasm.

Disagreement in the results was observed in CtpH and CtpI analysis using the different prediction tools. CtpI showed some hydrophobic amino acid sequences that could be considered part of TMS, but they do not fulfill all of the characteristics associated with TMS; meanwhile, CtpH did not show the expected hydrophobicity pattern for P-type ATPases (Figure 3). Therefore, at first it was difficult to determine whether CtpH and CtpI were in fact typical of P-type ATPases. Recent studies have classified CtpH and CtpI proteins of *M. bovis* as FUPA 24 (TC No. 3.A.3.24), i.e., "Functionally uncharacterized P-type ATPase family 24", and classify them as transporter proteins in the *TCDB-Transporter Classification Database* (<http://www.tcdb.org>). FUPA 24 proteins are homologous to P-type ATPases but have an unusually large N-terminal segment that makes them twofold longer than typical P-type ATPases [23] with two TMS. In addition, functional motifs of P-type ATPases are located within the C-terminal region of FUPA 24 [23].

By contrast, topology analysis in this study showed that *M. tuberculosis* H37Rv CtpH and CtpI contain more than the expected TMS (from 3 to 12) for FUPA 24 proteins. To overcome the discrepancy between TCDB and the six topology tools used in this work, the *TMDet* and *PPM* servers were used. These tools allow for the identification of TMS based on the tertiary structure of the protein; thus, not only can the correct size and composition of α -helices be guaranteed, but also their adequate disposition and organization in the plasma membrane can be determined. Ten TMS were found in CtpH and CtpI using both the *TMDet* and *PPM* servers. These segments are located with high probability in a hypothetical lipid bilayer as shown in Figure 4. This result strongly suggests that CtpH and CtpI have a type II topology, as has been determined for other members of the AEM group.

Hydrophobicity profiles suggest insights into ion transport specificity

The "topological clustering" based on the hydrophobicity profiles obtained with the *TMHMM 2.0* algorithm can be used as a predictive tool to determine substrate specificity for HM pumps [27]. Here, this strategy was applied to analyze HM, AEM and KdpB groups of *M. tuberculosis* H37Rv P-type ATPases. The hydrophobicity profiles from each *M. tuberculosis* P-type ATPase was compared with the obtained profiles from previously characterized P-type ATPases. As shown in Figure 3, the hydrophobicity profiles of CtpA, CtpB and CtpV were similar to the consensus Cu^+ P-type ATPases; the CtpC profile was similar to the different Zn^{2+} transporters Zn^{2+} , Cd^{2+}/Zn^{2+} and $Pb^{2+}/Cd^{2+}/Zn^{2+}/Hg^{2+}$. In addition, CtpD, CtpG and CtpJ had regions similar to the $Cd^{2+}/Zn^{2+}/Co^{2+}$ transporter. In the case of the AEM group, CtpF possesses TMS that were similar to the hydrophobic profiles obtained for consensus of Na^+ , Mg^{2+} and Na^+/K^+ ATPases. Additionally, CtpE was similar to Na^+ or H^+ transporters; however, its hydrophobicity profile is closer to that of Na^+ P-type ATPase. Considering the C-terminus region of CtpH (819 amino acid residues), this pump had a hydrophobicity profile very similar to the Ca^{2+} PMCA consensus. Finally, KdpB had the exact same hydrophobic profile as that which was determined for *E. coli* KdpB.

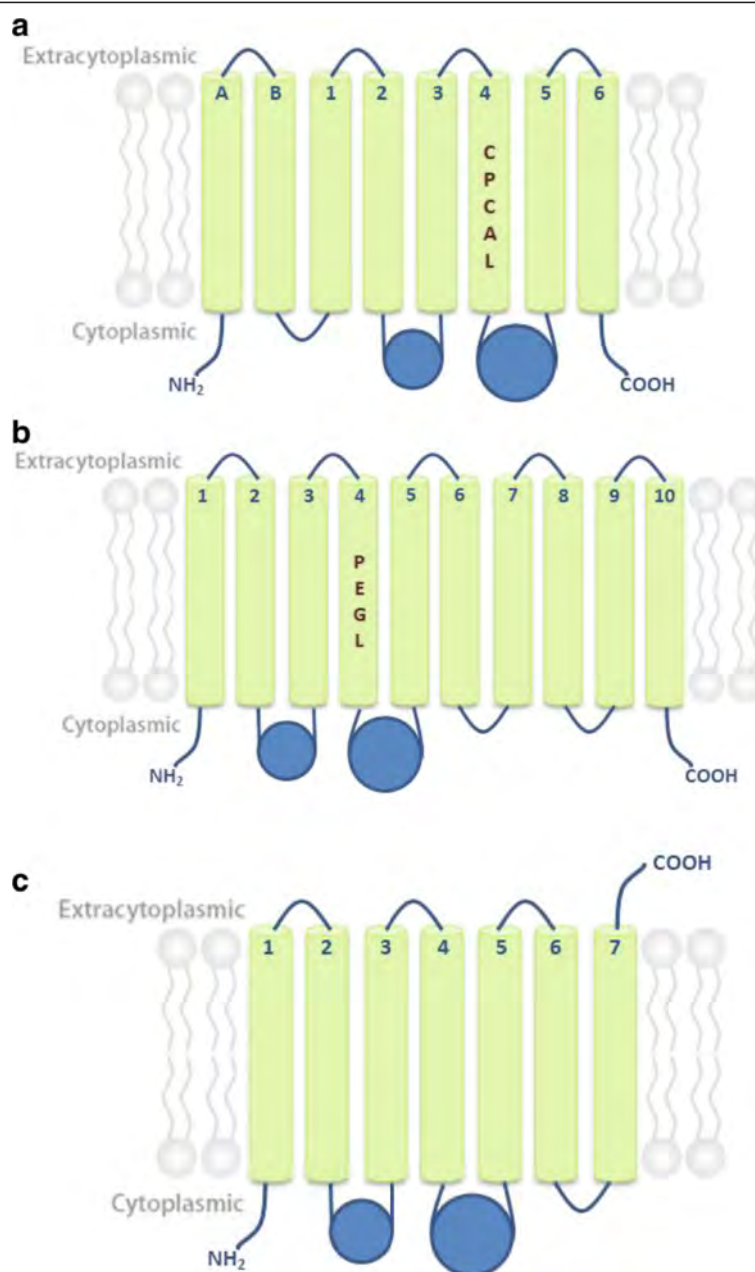


Figure 2 Predicted topology for P-type ATPases of MTBC. Type I topology (HM group ATPases) (a), type II topology (EAM group ATPases) (b), type III topology (KdpB group ATPases) (c), Ion binding motifs (number 4) are indicated in the corresponding transmembrane domain.

In general, there are the following two types of hydrophobicity profiles: the first comprises the AEM and the second that corresponds to HM. Interestingly, the relative position of TMS along amino acid sequences is highly conserved between both types of hydrophobicity profiles. The profile for AEM is characterized by a single TMS in the C-terminal region and two in the N-terminal region, whereas profiles for HM P-type ATPases have three highly hydrophobic regions, one of them in the N-terminus, one in the middle of the amino acid sequence, and the third in the C-terminal

end. Although the hydrophobicity profile of consensus sequences is similar among enzymes in each group, there are significant differences between profiles that allow their differentiation according to the transported ion.

P-type ATPases of the *Mycobacterium tuberculosis* complex possess the characteristic motifs involved in cation transport

The common catalytic mechanism of P-type ATPases is partially supported by conserved core sequences that are

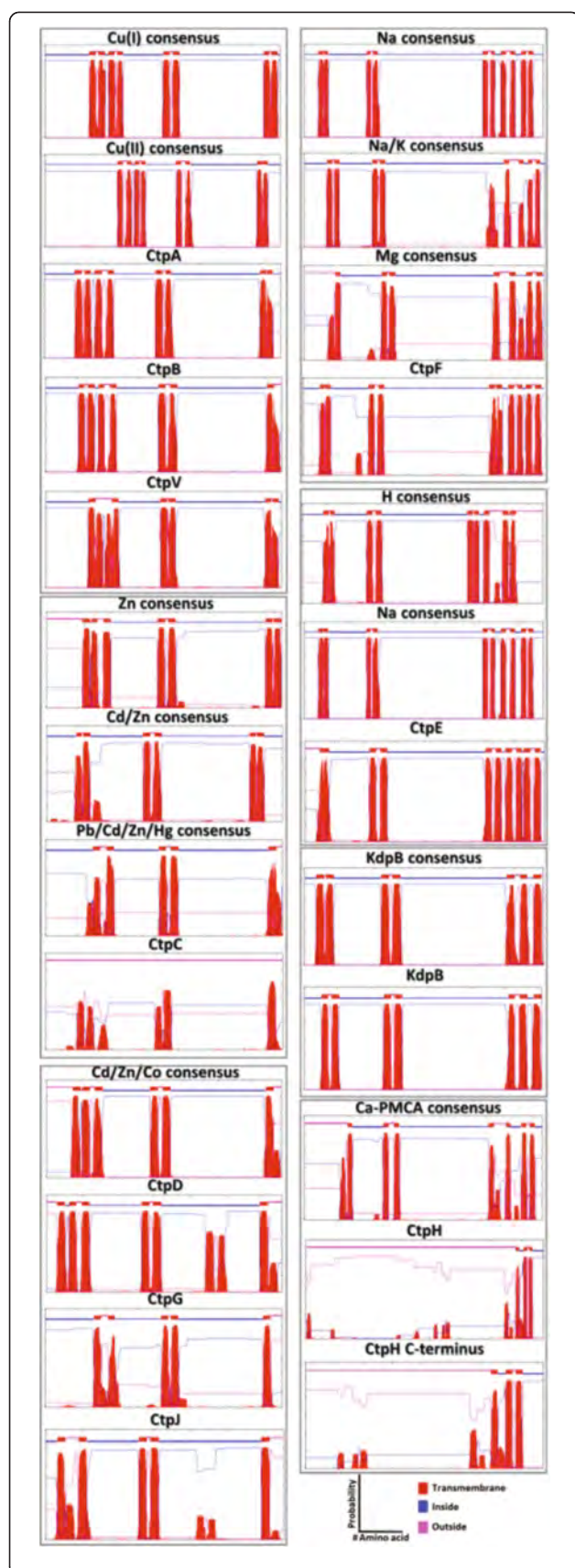


Figure 3 Hydrophobicity profiles of *M. tuberculosis* H37Rv P-type ATPases. Hydrophobicity profiles allow grouping of P-type ATPases with seven different ion specificities.

useful in the recognition of these pumps in proteomes [15]. To find the conserved motifs for P-type ATPases along the protein sequences, manual analysis was carried out because the servers commonly used for analysis of conserved motifs do not identify subtle variations within particular motifs of P-type ATPases. Nine conserved motifs typical for P-type ATPases were found in the MTBC pumps (Figure 5). We observed that each motif sequence was identical within orthologs of the MTBC. From N-terminus to the C-terminus, the motifs were as follows: motifs 1 [(PVA)G(DE)] and 2 [P(AS)D], related to conformational changes, together with motif 3 [TGE(SA)], associated with phosphatase activity, were located between TMS 2 and 3 in the actuator domain; the sequence of motif 4, which determines ion specificity, can be [PEG(LM)] or [(CSA)PCA(LV)], which was located at the end of the fourth TMS; the phosphorylation site [DKTGTLT] was found in the P domain; the remaining motifs, motif 6 [(KI)GA(PVA)(EDA)], an ATP binding facilitator, motifs 7 [(DV)(ASPI)(VP)(KAR)] and 8 [(MLV)I(TS)GD], involved in phosphorylation catalysts, and motif 9 [(VTC)AM(TV)GDG(VSAT)ND(AV)(PAL)A(LI)(RKA)(QMA)A(DNT)(VI)G(VI)(AG)(MV)], the hinge motif, which provides the flexibility necessary to achieve conformational changes during the pumping process, could be found between the fourth and fifth TMS [22]. Motifs 8 and 9 contain amino acid residues [TGDN and GDGXND] responsible for the coordination of Mg^{2+} as a cofactor of the enzymes [15,16,22,23].

It was observed that characteristic motifs of P-type ATPases exhibit slight variations compared with previously reported sequences for eukaryotic organisms [22]. The most conserved motif (motif 5) was almost the same within the 12 P-type ATPases, except CtpH and KdpB, which contained the conservative substitutions [DKTGTL(TS)] and [DKTGT(LI)T], respectively. The sequence of motif 4 [PEGL] has been previously associated with binding of AEM cations [22], and it has been found in CtpE, CtpF, and CtpI. Variation in the fourth position of this motif [PEGM] in CtpH has not been reported previously. CtpA, CtpB, CtpC, CtpD, CtpJ, CtpG and CtpV have the motif 4 sequence [(CSA)PCA(LV)] characteristic of HM transporters. Alternatively, KdpB, which corresponds to a β subunit of a multimeric P-type ATPase, did not have motif 4. This finding can be explained by the observation that this subunit mediates phosphorylation/dephosphorylation and

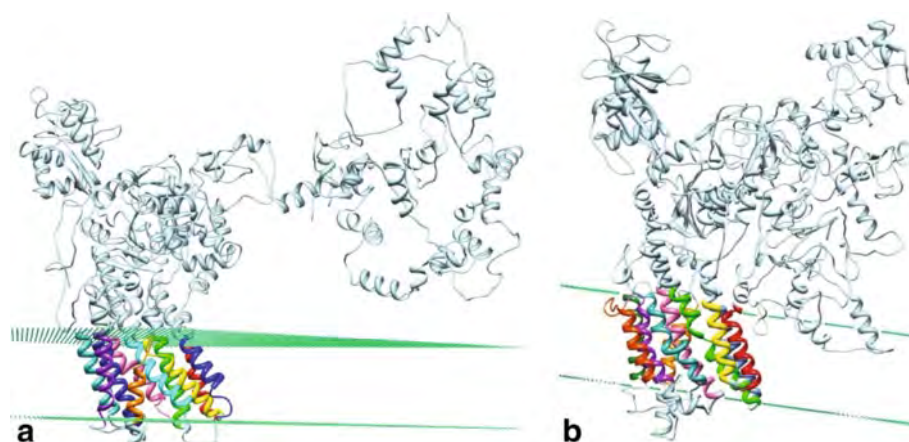


Figure 4 The CtpH and CtpI mycobacterial P-type ATPases (type II topology) exhibit ten TMS in their C-terminal half (different color helices). These tertiary structure models were generated with the *I-TASSER* tool and were modified with the *PPM* server [89] to include the possible location of the lipid bilayer. Dummies (green color) correspond to the location of the carbonyl groups in the bilayer. (a) CtpH model generated with the *PPM* server, (b) CtpI model built based on the *TMDet* results.

energy transduction during K^+ transport. In *E. coli*, the co-ordination function lies mainly in the KdpA subunit [26].

Cu^+ , Zn^{2+} and Co^{2+} pumping is mediated by P-type ATPases in the *M. tuberculosis* complex

The P_{IB} phylogenetic group of P-type ATPases is composed of HM pumps, and according to ion specificity, the P_{IB} group is subdivided into five subgroups (1 to 5). Alignments of consensus sequences from characterized P_{IB} ATPases with the HM P-type ATPases of MTBC were used for locating additional transmembrane motifs of this type of transporter [28-30], as shown in Figure 6. CtpA, CtpB and CtpV have the characteristic motifs of P_{IB-1} group members (Table 1), corroborating their ion transport specificity as predicted by hydrophobicity profiles, and suggesting that Cu^+ is transported preferentially to Cu^{2+} by these pumps.

This result is expected, because the special reducing conditions inside mycobacteria [27] are similar to the intra-phagosomal environment in macrophages where the bacilli reside. Recently, it was reported that the *M. tuberculosis* H37Rv CtpV is a Cu^+ exporter P-type ATPase, reinforcing our predictions [31]. CtpD and CtpJ exhibit the [SCP] and [HEGT] motifs of the P_{IB-4} group in the TMS6 and TMS8, respectively. Some reports have indicated that amino acids involved in Co^{2+} transport are still not well understood; for this reason, the absence of residue N in TMS7 cannot rule out the possibility that CtpD and CtpJ are Co^{2+} transporters. CtpC only exhibits the [CPC(X)₄S] motif in TMS6, which allows its classification as P_{IB-2} . In fact, it was recently reported as a Zn^{2+} P-type ATPase in *M. tuberculosis* H37Rv [25]; deficiency in this pump produced zinc accumulation within the mycobacterial cytoplasm, resulting in impaired intracellular

growth of tubercle bacilli [25]. Finally, CtpG does not have any of the additional motifs present in TMS6, TMS7 and TMS8; however, it possesses the [WI(YE)(RG)] sequence just before TMS6, between positions 406 and 409, and the [LS] motif located on TMS7, which is associated with Zn^{2+} P-type ATPases [32,33].

All of the results from this work provide predictive evidence for experimental studies to establish the ion specificity of MTBC P-type ATPases and their role in mycobacterial infection. It has been observed that some P-type ATPases of tubercle bacilli are over-expressed under conditions that mycobacteria face during infection [25,34,35]. For example, the expression level of CtpF and CtpC increase when *M. tuberculosis* is exposed to isoxyl, tetrahydrolipstatine and SRI#9190 antimicrobial compounds, indicating that these pumps might contribute to intrinsic resistance of mycobacteria to antimicrobial drugs [36]. Two additional studies indicate that hypoxia induces upregulation of the AEM transporter CtpF [37,38]; this observation is in agreement with those of another study that reported that CtpF is part of the regulon of the DosRS system, a relevant regulator of latency in *M. tuberculosis* [39]. In addition, upregulation of CtpF is observed *in vitro* when *M. tuberculosis* is incubated in the presence of S-nitroglutathione GSNO, ethanol, H_2O_2 and nitric oxide [40]. Moreover, CtpA, CtpC, CtpG, CtpV and CtpF are also induced when *M. tuberculosis* is phagocytized by macrophages [25,34,35].

Conclusion

Mycobacteria are unicellular organisms that respond to environmental stimuli, and the transport of substances across the plasma membrane could play a fundamental role in their adaptability. Computational analysis shows

	1	2	3	4	5	6	7	8	9
	PGD	PAD	TGES	PEGL	DKTGTLT	KGAPE	DPPR	MVTGD	VAVTGDGVNDSPALKKADIGVAM
CtpA	PDG	PAD	TGEA	CPCAL	DKTGTLT	AATVE	DAVK	LLTGD	VAMVGDGINDGPALARADLGMAI
Position	251	259	296	399	443	473	566	588	632
CtpB	PGE	AAD	TGEA	CPCAL	DKTGTLT	AAAVE	DTLK	LLTGD	VAMVGDGINDGPALVGADLGLAI
Position	276	281	299	402	446	480	575	597	641
CtpC	IGD	PVD	TGEN	CPCAV	DKTGTLT	KKASE	DEVK	MLTGD	VGMVGDGINDAPALAAADIGIAM
Position	238	251	269	364	408	505	538	561	605
CtpD	VGD	PAD	TGES	SPCAV	DKTGTLT	RGAPL	DQLR	LLTGD	VLLVGDGVNDAPAMAAARAAMVAM
Position	169	182	200	303	347	437	471	494	538
CtpE	PGD	VVD	TGEA	PEGL	DKTGTLT	IGAPD	DARE	VISGD	VAMTGDGVNDVLALKDADIGVAM
Position	129	134	153	258	301	386	447	464	531
CtpF	PGD	PAD	TGES	PEGL	DKTGTLT	KGAPE	DPPR	MITGD	VAMTGDGVNDAPALRQANIGVAM
Position	142	155	174	290	333	465	541	563	638
CtpG	VGD	ATD	TGES	APCAL	DKTGTLT	AAALE	DELK	MLTGD	TAMVGDGVNDAPALAAADLGIAM
Position	285	298	316	418	462	492	581	603	646
CtpH	PGD	PAD	TGES	PEGM	DKTGTLS	KGAPE	DTPR	LITGD	CAMVGDGSNDAAAIRAATVIGIV
Position	776	789	808	922	965	1060	1132	1154	1225
CtpI	VGD	PAD	TGES	PEGL	DKTGTLT	KGAPE	DTAR	LITGD	TAMVGDGANDAAAIRMADVIGIV
Position	865	878	897	1010	1053	1167	1243	1265	1335
CtpJ	IGD	SAD	TGEP	SPCAV	DKTGTLT	RGTPE	DQLR	LLTG D	LTVVGDGINDAPALAAAHVGIAM
Position	162	175	193	296	340	347	472	495	539
CtpV	VGD	PVD	TGES	CPCAL	DKTGTLT	AAAVE	DTVK	MITGD	VAMVGDGVNDAPALVQADLGLAI
Position	282	295	313	416	460	490	589	611	655
KdpB	AGE	SVD	TGES	-	DKTGTIT	KGAAA	DEMIR	MITGD	VAMTGDGTNDAPALAAQADVGVAM
Position	153	169	176	-	324	417	485	496	540
HM Consensus	VGD	PAD	TGES	CPCAL	DKTGTLT	AAAVE	DTLR	LLTGD	VAMVGDGINDAPALAAADLGIAM
H AM Consensus	PGD	PAD	TGES	PEGL	DKTGTLT	KGAPE	DTPR	LITGD	VAMTGDGVNDAAALRXADIGVAM

Figure 5 Conserved motifs for the P-type ATPases identified in *M. tuberculosis* H37Rv. The conserved motifs were compared with the motifs characteristic of the P-type ATPase superfamily. Identical residues, conserved substitutions, semi-conserved substitutions and unrelated residues are indicated in red, blue, green and black, respectively.

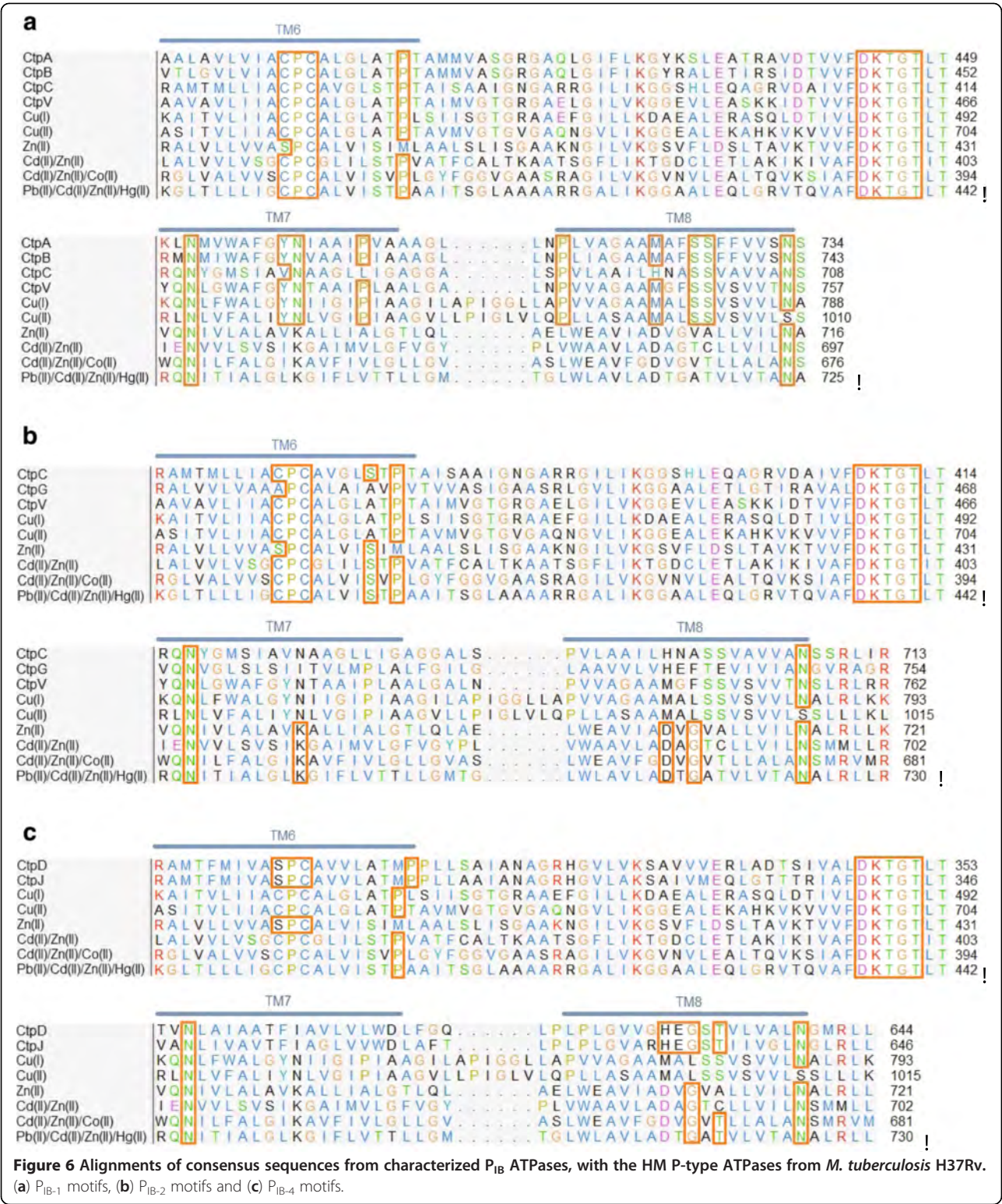


Table 1 Additional conserved motifs observed in heavy metal transporter P-type ATPases

Phylogenetic subgroup	Cation specificity	TMS6	TMS7	TMS8	MTBC pump
P _{IB-1}	Cu ⁺	CPC	YN	MXXSS	CtpA
					CtpB
					CtpV
P _{IB-2}	Zn ²⁺	CPC(X) ₄ S	K	DG	CtpC *
P _{IB-3}	Cu ²⁺	CPH	YN	MXXS	-
P _{IB-4}	Co ²⁺	SCP	N	HEGT	CtpD *
					CtpJ *
P _{IB-5}	Undetermined	TPCP	QX ₄ GX ₃ SX ₃ M	PX ₅ QEX ₂ DX ₅ N	-

The additional motifs are located on TMS6, TMS7 and TMS8. *, pumps without a complete set of motifs (see discussion).

that each MTBC species has a consistent aggrupation of the 12 P-type ATPases involved in ion transport. In this context, *M. tuberculosis* strains H37Ra and H37Rv share identical sequences for P-type ATPases, facilitating subsequent genetic studies using the attenuated strain H37Ra. The large number of HM P-type ATPases expressed by the MTBC strongly suggests that they could be essential for the bacteria to counteract the increased level of HM accumulated by macrophages after infection with tubercle bacilli. Thus, compensatory ion transport strategies could be used by mycobacteria to survive in host cells.

The different bioinformatics approaches used in this work to analyze the P-type ATPases identified in the MTBC are in agreement with the initial classification from the HMM search. The results obtained show that *M. tuberculosis* has the following three groups of P-type ATPases: HM transporters (CtpA, CtpB, CtpC, CtpD, CtpG, CtpJ and CtpV), AEM transporters (CtpE, CtpF, CtpH, and CtpI) and the KdpB protein, which corresponds to the β subunit of a multimeric K⁺ ATPase transporter exclusive to prokaryotes. Hydrophobicity analysis identified α -helix type TMS grouped into the following three topological types: type I (HM group), type II (AEM group) and type III (KdpB group). Interestingly, we report a possible mis-annotation for CtpH and CtpI in the TCDB Database, where they are classified as FUPA 24 type with two TMS, unlike the ten TMS identified for these unusually large transporters in this work. Finally, a counterpart of non-catalytic β subunits of Na⁺/K⁺ or H⁺/K⁺ ATPases does not exist within the MTBC proteomes.

Competing interest

There are no conflicts of interest to declare.

Authors' contributions

LN-A, AL-T, MP-R, JC-B, L-M S, DL, LM-R and CY-S wrote the manuscript, validated the tools and carried out the data analysis and interpretation. LN-A, AL-T, MP-R and JC-B contributed to the methodological design, supervised its development and critically revised the manuscript's content. All authors read and approved the final version of the manuscript.

Acknowledgements

This work was supported by the Dirección de Investigación Bogotá (DIB)-Universidad Nacional de Colombia, grants 12351, 12176, 11875, 11882 and Colciencias grant 13802. This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine and National Center for Biotechnology Information.

Funding

This work was supported by the Dirección de Investigación Bogotá (DIB)-Universidad Nacional de Colombia, grants 12351, 12176, 11875, 11882 and Colciencias grant 13802. This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine and National Center for Biotechnology Information. The funding sources had no involvement in the study design; collection, analysis and interpretation of data; the writing of the manuscript or the decision to submit the article for publication.

Author details

¹Chemistry Department, Faculty of Sciences, Universidad Nacional de Colombia, Bogotá, Colombia, Carrera 30 # 45-03, Ciudad Universitaria, Bogotá, Colombia. ²Computational Biology Branch, NCBI, NLM, NIH, Bethesda, USA. ³PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia.

Received: 25 June 2012 Accepted: 27 September 2012

Published: 3 October 2012

References

- WHO: *Global tuberculosis control: WHO report 2011*. Switzerland: Publications of the World Health Organization; 2011:246.
- McEvoy CR, et al: *The role of IS6110 in the evolution of Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* 2007, **87**(5):393-404.
- Issa R, et al: *Detection and discrimination of Mycobacterium tuberculosis complex*. *Diagn Microbiol Infect Dis* 2012, **72**(1):62-7.
- Knechel NA: *Tuberculosis: pathophysiology, clinical features, and diagnosis*. *Crit Care Nurse* 2009, **29**(2):34-43. quiz 44.
- Tufariello JM, Chan J, Flynn JL: *Latent tuberculosis: mechanisms of host and bacillus that contribute to persistent infection*. *Lancet Infect Dis* 2003, **3**(9):578-90.
- Koul A, et al: *The challenge of new drug discovery for tuberculosis*. *Nature* 2011, **469**(7331):483-90.
- Nagata T, et al: *Comparative molecular biological analysis of membrane transport genes in organisms*. *Plant Mol Biol* 2008, **66**(6):565-85.
- Axelsen KB, Palmgren MG: *Evolution of substrate specificities in the P-type ATPase superfamily*. *J Mol Evol* 1998, **46**(1):84-101.
- Pedersen PL: *Transport ATPases into the year 2008: a brief overview related to types, structures, functions and roles in health and disease*. *J Bioenerg Biomembr* 2007, **39**(5-6):349-55.
- Rocafull MA, et al: *Isolation and cloning of the K⁺-independent, ouabain-insensitive Na⁺-ATPase*. *Biochim Biophys Acta* 2011, **1808**(6):1684-700.

11. Pinoni SAL, A A: **Na⁺ ATPase activities in chela muscle of the euryhaline crab *Neohelice granulata*: differential response to environmental salinity.** *J Exp Mar Bio Ecol* 2009, **372**(1-2):91-97.
12. Jorgensen PL: **Purification and characterization of (Na⁺, K⁺)-ATPase. V. Conformational changes in the enzyme Transitions between the Na-form and the K-form studied with tryptic digestion as a tool.** *Biochim Biophys Acta* 1975, **401**(3):399-415.
13. Jorgensen PL, Hakansson KO, Karlsh SJ: **Structure and mechanism of Na, K-ATPase: functional sites and their interactions.** *Annu Rev Physiol* 2003, **65**:817-49.
14. Xu C, et al: **A structural model for the catalytic cycle of Ca(2+)-ATPase.** *J Mol Biol* 2002, **316**(1):201-11.
15. Palmgren MG, Nissen P: **P-type ATPases.** *Annu Rev Biophys* 2011, **40**:243-66.
16. Kuhlbrandt W: **Biology, structure and mechanism of P-type ATPases.** *Nat Rev Mol Cell Biol* 2004, **5**(4):282-95.
17. Bublitz M, et al: **In and out of the cation pumps: P-type ATPase structure revisited.** *Curr Opin Struct Biol* 2010, **20**(4):431-9.
18. Uniprot. cited 2011; Available from: <http://www.uniprot.org/>.
19. Pirovano W, Feenstra KA, Heringa J: **PRALINETM: a strategy for improved multiple alignment of transmembrane proteins.** *Bioinformatics* 2008, **24**(4):492-7.
20. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**(9):755-63.
21. Zhang Y: **I-TASSER server for protein 3D structure prediction.** *BMC Bioinforma* 2008, **9**:40.
22. Thever MD, Saier MH Jr: **Bioinformatic characterization of P-type ATPases encoded within the fully sequenced genomes of 26 eukaryotes.** *J Membr Biol* 2009, **229**(3):115-30.
23. Chan H, et al: **The P-type ATPase superfamily.** *J Mol Microbiol Biotechnol* 2010, **19**(1-2):5-104.
24. Cole ST: **Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence.** *Nature* 1998, **393**:537.
25. Botella H, et al: ***Mycobacterial* p(1)-type ATPases mediate resistance to zinc poisoning in human macrophages.** *Cell Host Microbe* 2011, **10**(3):248-59.
26. Bramkamp M, Altendorf K, Greie JC: **Common patterns and unique features of P-type ATPases: a comparative view on the KdpFABC complex from *Escherichia coli* (Review).** *Mol Membr Biol* 2007, **24**(5-6):375-86.
27. Lewinson O, Lee AT, Rees DC: **A P-type ATPase importer that discriminates between essential and toxic transition metals.** *Proc Natl Acad Sci U S A* 2009, **106**(12):4677-82.
28. Argüello JM: **Identification of Ion-Selectivity Determinants in Heavy-Metal Transport P1B-type ATPases.** *J Membr Biol* 2003, **195**:93-108.
29. Argüello JM, Eren E, González-Guerrero M: **The structure and function of heavy metal transport P1B ATPases.** *Biometals* 2007, **20**:233-248.
30. Argüello JM, Gonzalez-Guerrero M, Raimunda D: **Bacterial transition metal P(1B)-ATPases: transport mechanism and roles in virulence.** *Biochemistry* 2011, **50**(46):9940-9.
31. Ward SK, et al: **CtpV: a putative copper exporter required for full virulence of *Mycobacterium tuberculosis*.** *Mol Microbiol* 2010, **77**(5):1096-1110.
32. Lewinson O, Lee AT, Rees DC: **A P-type ATPase importer that discriminates between essential and toxic transition metals.** *Proc Natl Acad Sci* 2009, **106**(12):4677-4682.
33. Futai M, Wada Y, Kaplan JH: *Handbook of ATPases: Biochemistry, Cell Biology, Pathophysiology.* 1 ed. Weinheim: WILEY-VCH Verlag GmbH & Co. KGaA; 2004.
34. Schnappinger D, et al: **Transcriptional Adaptation of *Mycobacterium tuberculosis* within Macrophages: Insights into the Phagosomal Environment.** *J Exp Med* 2003, **198**(5):693-704.
35. Kumar M, et al: **Identification of *Mycobacterium tuberculosis* genes preferentially expressed during human infection.** *Microb Pathog* 2011, **50**(1):31-8.
36. Waddell SJ, et al: **The use of microarray analysis to determine the gene expression profiles of *Mycobacterium tuberculosis* in response to anti-bacterial compounds.** *Tuberculosis (Edinb)* 2004, **84**(3-4):263-74.
37. Bacon J, et al: **The influence of reduced oxygen availability on pathogenicity and gene expression in *Mycobacterium tuberculosis*.** *Tuberculosis (Edinb)* 2004, **84**(3-4):205-17.
38. Sherman DR, et al: **Regulation of the *Mycobacterium tuberculosis* hypoxic response gene encoding alpha -crystallin.** *Proc Natl Acad Sci U S A* 2001, **98**(13):7534-9.
39. Kendall SL, et al: **The *Mycobacterium tuberculosis* dosRS two-component system is induced by multiple stresses.** *Tuberculosis (Edinb)* 2004, **84**(3-4):247-55.
40. Voskuil MI, et al: **Inhibition of respiration by nitric oxide induces a *Mycobacterium tuberculosis* dormancy program.** *J Exp Med* 2003, **198**(5):705-13.

doi:10.1186/1472-6807-12-25

Cite this article as: Novoa-Aponte et al.: *In silico* identification and characterization of the ion transport specificity for P-type ATPases in the *Mycobacterium tuberculosis* complex. *BMC Structural Biology* 2012 **12**:25.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



Genome Sequences for Six *Rhodanobacter* Strains, Isolated from Soils and the Terrestrial Subsurface, with Variable Denitrification Capabilities

Joel E. Kostka,^{a,f} Stefan J. Green,^b Lavanya Rishishwar,^a Om Prakash,^c Lee S. Katz,^d Leonardo Mariño-Ramírez,^{e,f} I. King Jordan,^{a,f} Christine Munk,^g Natalia Ivanova,^g Natalia Mikhailova,^g David B. Watson,^h Steven D. Brown,ⁱ Anthony V. Palumbo,ⁱ and Scott C. Brooks^h

School of Biology, Georgia Institute of Technology, Atlanta, Georgia, USA^a; DNA Services Facility, Research Resource Center, University of Illinois, Chicago, Illinois, USA^b; National Centre for Cell Science, Pune, India^c; Centers for Disease Control and Prevention, Atlanta, Georgia, USA^d; National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland, USA^e; PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia^f; United States Department of Energy Joint Genome Institute, Walnut Creek, California, USA^g; Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA^h; and Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USAⁱ

We report the first genome sequences for six strains of *Rhodanobacter* species isolated from a variety of soil and subsurface environments. Three of these strains are capable of complete denitrification and three others are not. However, all six strains contain most of the genes required for the respiration of nitrate to gaseous nitrogen. The nondenitrifying members of the genus lack only the gene for nitrate reduction, the first step in the full denitrification pathway. The data suggest that the environmental role of bacteria from the genus *Rhodanobacter* should be reevaluated.

The genus *Rhodanobacter* contains 11 described species of Gram-negative, non-spore-forming, rod-shaped bacteria belonging to the family *Xanthomonadaceae* and the class *Gamma-proteobacteria* of the phylum *Proteobacteria*. Described species have been isolated mainly under aerobic conditions from surficial soils (1, 4, 5, 9, 12, 15, 16). Denitrification has not been considered a property of this genus. Recently, two strains of a new species, *Rhodanobacter denitrificans*, were isolated from a contaminated terrestrial subsurface environment and shown to denitrify (7, 13). Furthermore, nitrate-reducing isolates were recently recovered from sewage sludge (17), and we and others determined that *Rhodanobacter thiooxydans* is capable of denitrification (13, 14). In some acidic and nitrate-rich environments, *Rhodanobacter* species dominate bacterial communities (8, 14).

To explore the genetic basis of phenotypes leading to bacterial community dominance in such environments, genome sequences were acquired for three denitrifying strains (*R. denitrificans* 2APBS1 and 116-2 and *R. thiooxydans*) and three strains incapable of denitrification (*Rhodanobacter fulvus*, *Rhodanobacter spathiphylli*, and *Rhodanobacter* sp. 115). A complete *R. denitrificans* 2APBS1^T genome sequence was generated using paired-end Illumina and Roche 454 mate-pair sequencing and manual finishing steps, essentially as described previously (3, 6). Four draft genomes (*R. denitrificans* 116-2, *R. thiooxydans*, *R. fulvus*, and *R. spathiphylli*) were generated by *de novo* assembly of paired-end Illumina sequence data (~5.7 to 9.5 million paired-end reads/genome, yielding ~1.1 to 1.9 Gb of total output/genome) (CLC Genomics Workbench 5.0; CLC bio A/S, Denmark). DNA from each strain was prepared for sequencing using the Nextera library preparation kit (Epicentre, Madison, WI). DNA from *Rhodanobacter* sp. 115 was prepared for sequencing using the Ion Xpress fragment library kit (Life Technologies, Grand Island, NY) and sequenced using a Personal Genome Machine (Ion Torrent, San Francisco, CA), yielding approximately 1.4 Mb of reads (~138 Mb of total output). For *Rhodanobacter* sp. 115, genome assembly was performed as described previously (10) using CG-Pipeline modules (11), yielding 453 contigs and 4.2 Mb of genomic sequence data.

The complete genome of *R. denitrificans* 2APBS1 is 4.23 Mb. Annotation was performed in RAST (2) and in the CG-Pipeline before being submitted to NCBI.

Denitrification is a strain-specific trait, and the high sequence divergence observed in genetic markers for denitrification challenges our ability to understand the fundamental ecological principles and environmental parameters controlling nitrate attenuation in terrestrial environments (7). Thus, whole-genome sequencing of closely related denitrifying and nondenitrifying taxa is essential to improve detection of denitrifying bacteria in the environment and to develop hypotheses regarding the distribution and acquisition of denitrification genes. Comparative analysis of the six genomes revealed that all strains contained genes coding for complete or nearly complete denitrification pathways. The three nondenitrifying lineages lacked only genes for nitrate reduction. These organisms may still be capable of denitrification, however. Nitrate to nitrite reduction is a widespread physiological capability in the bacterial domain, and in complex environments, such as soil, nitrite will be available for organisms capable of nitrite reduction to gaseous nitrogen end products. These data indicate that the environmental role of bacteria from the genus *Rhodanobacter* should be reevaluated.

Nucleotide sequence accession numbers. The *Rhodanobacter* genome assemblies and their annotations were deposited in GenBank under the accession numbers [AGIL00000000](#) (*Rhodanobacter* strain 116-2, DSM 23569), [AJXS00000000](#) (*Rhodanobacter* strain 115), [AJXT00000000](#) (*Rhodanobacter* strain 116-2, DSM 17631), [AJXU00000000](#) (*Rhodanobacter* strain 115, DSM 18449), [AJXV00000000](#) (*Rhodanobacter* strain 115, DSM 24678), and [AJXW00000000](#) (*Rhodanobacter* strain 115, DSM 18863).

Received 17 May 2012 Accepted 4 June 2012

Address correspondence to Joel E. Kostka, joel.kostka@biology.gatech.edu.

Copyright © 2012, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JB.00871-12

ACKNOWLEDGMENTS

This research was supported by the Office of Science (BER), U.S. Department of Energy, grant numbers DE-FG02-07ER64373, -97ER62469, and -97ER64398 and by the Oak Ridge Integrated Field-Research Challenge, operated by the Environmental Sciences Division, Oak Ridge National Laboratory (ORNL).

ORNL is managed by UT-Battelle, LLC, for the U.S. Department of Energy contract no. DE-AC05-00OR22725.

This research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI.

The complete genome of *Rhodanobacter denitrificans* strain 2APBS1 was sequenced by the U.S. Department of Energy Joint Genome Institute, supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231.

We gratefully acknowledge assistance from Tonia Mehlhorn and Kenneth Lowe for sampling and uranium measurements.

REFERENCES

1. An DS, Lee HG, Lee ST, Im WT. 2009. *Rhodanobacter ginsenosidimutans* sp. nov., isolated from soil of a ginseng field in South Korea. *Int. J. Syst. Evol. Microbiol.* 59:691–694.
2. Aziz RK, et al. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75.
3. Bennett, S. 2004. Solexa Ltd. *Pharmacogenomics* 5:433–438.
4. Bui TP, Kim YJ, Kim H, Yang DC. 2010. *Rhodanobacter soli* sp. nov., isolated from soil of a ginseng field. *Int. J. Syst. Evol. Microbiol.* 60:2935–2939.
5. De Clercq D, et al. 2006. *Rhodanobacter spathiphylli* sp. nov., a gamma-proteobacterium isolated from the roots of *Spathiphyllum* plants grown in a compost-amended potting mix. *Int. J. Syst. Evol. Microbiol.* 56:1755–1759.
6. Elkins JG, et al. 2010. Complete genome sequence of the cellulolytic thermophile *Caldicellulosiruptor obsidiansis* OB47T. *J. Bacteriol.* 192:6099–6100.
7. Green SJ, et al. 2010. Denitrifying bacteria isolated from terrestrial subsurface sediments exposed to mixed-waste contamination. *Appl. Environ. Microbiol.* 76:3244–3254.
8. Green SJ, et al. 2012. Denitrifying bacteria from the genus *Rhodanobacter* dominate bacterial communities in the highly contaminated subsurface of a nuclear legacy waste site. *Appl. Environ. Microbiol.* 78:1039–1047.
9. Im WT, Lee ST, Yokota A. 2004. *Rhodanobacter fulvus* sp. nov., a β -galactosidase-producing gammaproteobacterium. *J. Gen. Appl. Microbiol.* 50:143–147.
10. Jordan IK, et al. 2011. Genome sequences for five strains of the emerging pathogen *Haemophilus haemolyticus*. *J. Bacteriol.* 193:5879–5880.
11. Kislyuk AO, et al. 2010. A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics* 26:1819–1826.
12. Nalin R, Simonet P, Vogel TM, Normand P. 1999. *Rhodanobacter lindaniclasticus* gen. nov., sp. nov., a lindane-degrading bacterium. *Int. J. Syst. Bacteriol.* 49:19–23.
13. Prakash O, et al. 25 November 2011, posting date. Description of *Rhodanobacter denitrificans* sp. nov., isolated from uranium and nitrate contaminated subsurface sediment. *Int. J. Syst. Evol. Microbiol.* doi:10.1099/ijs.0.035840-0.
14. van den Heuvel RN, van der Biezen E, Jetten MSM, Hefting MM, Kartal B. 2010. Denitrification at pH 4 by a soil-derived *Rhodanobacter*-dominated community. *Environ. Microbiol.* 12:3264–3271.
15. Wang L, et al. 2011. *Rhodanobacter panaciterrae* sp. nov., a bacterium with ginsenoside-converting activity isolated from soil of a ginseng field. *Int. J. Syst. Evol. Microbiol.* 61:3028–3032.
16. Weon HY, et al. 2007. *Rhodanobacter ginsengisoli* sp. nov. and *Rhodanobacter terrae* sp. nov., isolated from soil cultivated with Korean ginseng. *Int. J. Syst. Evol. Microbiol.* 57:2810–2813.
17. Woo SG, Srinivasan S, Kim MK, Lee M. 6 January 2012, posting date. *Rhodanobacter caeni* sp. nov., a denitrifying bacterium isolated from sludge in a sewage disposal plant. *Int. J. Syst. Evol. Microbiol.* doi:10.1099/ijs.0.033365-0.



Differences in local genomic context of bound and unbound motifs

Loren Hansen ^{a,b}, Leonardo Mariño-Ramírez ^{a,c,*}, David Landsman ^a

^a Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8900 Rockville Pike, Bethesda, MD 20894, USA

^b Bioinformatics Program, Boston University, Boston, MA 02215, USA

^c PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia

ARTICLE INFO

Article history:

Accepted 4 June 2012

Available online 10 June 2012

Keywords:

Gene regulation

Yeast

Transcription factors

Genomic features

Machine learning

ABSTRACT

Understanding gene regulation is a major objective in molecular biology research. Frequently, transcription is driven by transcription factors (TFs) that bind to specific DNA sequences. These motifs are usually short and degenerate, rendering the likelihood of multiple copies occurring throughout the genome due to random chance as high. Despite this, TFs only bind to a small subset of sites, thus prompting our investigation into the differences between motifs that are bound by TFs and those that remain unbound. Here we constructed vectors representing various chromatin- and sequence-based features for a published set of bound and unbound motifs representing nine TFs in the budding yeast *Saccharomyces cerevisiae*. Using a machine learning approach, we identified a set of features that can be used to discriminate between bound and unbound motifs. We also discovered that some TFs bind most or all of their strong motifs in intergenic regions. Our data demonstrate that local sequence context can be strikingly different around motifs that are bound compared to motifs that are unbound. We concluded that there are multiple combinations of genomic features that characterize bound or unbound motifs.

Published by Elsevier B.V.

1. Introduction

Control of gene expression is fundamental to all forms of life. Transcription initiation is controlled primarily by transcription factor (TF) binding to key DNA sequence motifs. In many cases, the sequence motifs recognized by DNA binding proteins are short and degenerate, thus rendering it highly likely that they may appear multiple times in the genome due to random chance. This is especially true for large eukaryotic genomes. Using sequence motifs alone to predict TF binding leads to an unacceptable level of false positives (D'Haeseleer, 2006; Fickett, 1996). Given this, many transcription factor binding site prediction methods incorporate other sources of information in addition to sequence similarity. For example, previous studies have incorporated information about the sequence conservation between species (Blanchette and Tompa, 2003; Xie et al., 2005) or the fact that different TFs frequently bind DNA in clusters (Frith et al., 2003; Tharakaraman et al., 2008). Combining sequence conservation or clustering of TFBS into a single tool can improve predictive performance; however how a TF selects the appropriate motif out of all its occurrences *in vivo* across the

genome remains unknown. While conservation between species is a useful computational tool for identifying potential regulatory regions, this information is not available *in vivo* to guide a TF to the correct binding location. Recent advances in high-throughput techniques [e.g., chromatin immunoprecipitation with microarray technology (ChIP-chip) and chromatin immunoprecipitation sequencing (ChIP-seq)] have produced high-quality maps of genome-wide TF binding. In addition, machine-learning techniques have been used successfully to predict TF binding (Bauer et al., 2010; Holloway et al., 2005, 2007). In this study, we used both these techniques to compare the local genomic environment near established TF binding sites with unbound motifs to identify biological features associated with bound motifs *in vivo*. Our approach is to use machine-learning techniques not primarily in an attempt to predict TF binding sites but rather to gain insight into local genomic features of bound and unbound motifs.

2. Materials and methods

2.1. Data sets

Histone modification data was obtained from a previous study (Pokholok et al., 2005). We obtained the raw data from the ArrayExpress Archive (<http://www.ebi.ac.uk/arrayexpress/>) and performed MA2C normalization (Peng et al., 2007). Nucleosome occupancy data was previously published (Kaplan et al., 2009), and the nucleosome occupancy scores as calculated by the authors were used unchanged. PWMs used in this study were obtained from three different sources. The matrices used to produce the set of bound and

Abbreviations: TF, Transcription factors; DNA, Deoxyribonucleic acid; ChIP-chip, Chromatin immunoprecipitation with microarray technology; ChIP-Seq, Chromatin immunoprecipitation sequencing; PWM, Position weight matrix; TSS, Transcription start site.

* Corresponding author at: Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8900 Rockville Pike, Bethesda, MD 20894, USA. Tel.: +1 301 402 3708; fax: +1 301 480 2288.

E-mail address: mario@ncbi.nlm.nih.gov (L. Mariño-Ramírez).

unbound motifs were taken from Maclsaac et al. (2006). This was performed in an attempt to be consistent with the matrices that were used by Maclsaac et al. to produce a map of TF binding across the yeast genome. We used 143 matrices from Badis et al. (2008) and Zhu et al. (2009) to identify motifs near bound or unbound motifs. Many of the DNA specificity matrices supplied by the authors were position frequency matrices, which we converted to PWMs as described previously (Wasserman and Sandelin, 2004). Gene coordinates were obtained from the UCSC genome browser (<http://genome.ucsc.edu/>).

2.2. Construction of bound and unbound motif datasets

We obtained the binding locations within *Saccharomyces cerevisiae* intergenic regions of 118 TFs that were mapped using ChIP-chip (Maclsaac et al., 2006). Maclsaac et al. analyzed a previously published ChIP-chip dataset (Harbison et al., 2004). Selected for further study were those TFs with at least 100 experimentally mapped binding sites ($n=12$). We used the mapped binding sites generated under the strictest criteria as defined by Maclsaac et al. Briefly, sites defined as bound required a PWM match of 60% of the maximum possible log-likelihood PWM score, conservation in at least three of four *sensu stricto* yeast species, and a p-value of less than 0.001 for the probe containing the motif. A recent study using ChIP-exo argues the false positive rate for ChIP-chip may be as high as 50% (Rhee and Pugh, 2011). By using both motif information and conservation between species Maclsaac et al. were able to identify bound motifs with high confidence.

The bound sites as obtained by Maclsaac et al. are presumably centered at motif occurrences. But the bound sites as listed do not have information we would like to study such as the average strength of the bound motifs. Hence it was necessary to remap motif locations and link motifs with bound sites as identified by Maclsaac et al. To do so we scanned yeast intergenic regions with the PWMs corresponding to the 12 selected TFs and all occurrences of motifs with a score of 60% or greater maximum log-likelihood score were collected. Motifs that overlapped the experimentally mapped binding sites as defined by Maclsaac et al. were labeled as “bound” motifs with the restriction of only one bound motif allowed per experimentally mapped site. Given that bound sites as defined by Maclsaac et al. must include a motif instance, it seems that all experimentally determined bound sites taken from the Maclsaac study should overlap a motif as identified by us. In practice this was largely the case (Supplemental Table 1) with one exception, DIG1. Other experimentally mapped TF binding sites as defined by Maclsaac et al. also do not exhibit complete overlap with a motif as defined by us (Supplemental Table 1 second column). This is likely due to Maclsaac using an older yeast genome build. The 11 TFs retained for further analysis were: REB1, GCN4, MBP1, PHD1, SKN7, STE12, SUT1, SWI4, SWI6, ABF1, and CBF1. Only binding sites as determined by Maclsaac et al. for which we could link a motif were defined as “bound” motifs.

To produce the set of unbound motifs, we obtained ChIP-chip binding data for the 11 TFs used in this study (Harbison et al., 2004). A p-value of binding was assigned to each intergenic region in yeast according to Harbison et al. A set of unbound motifs was produced by scanning yeast intergenic regions with PWMs corresponding to the 11 TFs; all occurrences of motifs with a 60% or greater maximum log-likelihood score were identified. Motifs found in intergenic regions whose p-value of binding was 0.5 or greater in all experimental conditions studied by Harbison et al. were labeled as unbound.

To check the robustness of our approach we repeated the analysis shown in Figs. 2, 3 and 4 using p-value cutoffs of 0.4 and 0.6 in calling unbound motifs, our results did not change.

2.3. Generation of feature vectors

2.3.1. Generation of nucleosome-based features

Pokholok et al. (2005) used tiling arrays to map histone modifications in *S. cerevisiae*. We used this data to calculate the level of histone

modification around bound and unbound motifs. For each 200-bp window centered on a motif, we obtained the degree of enrichment by averaging the normalized log ratio values of the probes within that region. For example, the feature “H3K14ac” represents the average degree of acetylation of lysine 14 in histone H3 for the given window. A similar approach was used for each histone modification mark.

To calculate the degree of nucleosome occupancy, we used a dataset produced by Kaplan et al. (2009). For most positions in the genome, Kaplan and co-authors calculated a nucleosome occupancy score. The average nucleosome occupancy was normalized to zero. A value greater than zero represented nucleosome enrichment relative to the genome-wide average, while a value less than zero signified nucleosome depletion. For each window centered at a motif, nucleosome occupancy was calculated by averaging the nucleosome occupancy scores for that window. Eight features were chromatin-based: “Nucleosome occupancy,” “H3K14ac,” “H3K36me3,” “H3K4me1,” “H3K4me2,” “H3K4me3,” “H3K79me3,” and “H3K9ac.”

2.3.2. Generation of motif-based features

We scripted our own program in Perl to scan yeast intergenic regions with a library of 143 PWMs obtained as described above. Motif matches of 70% or better of the maximum possible log-likelihood score for the given PWM were retained. For each bound and unbound motif for the set of nine TFs analyzed by feature selection, the number of motif matches within 100 bp of every motif represented in our PWM library was calculated. We did not consider any motif match that was within 10 bp of the bound or unbound motifs. 143 out of the 171 features were generated in this fashion. An additional motif-based feature was motif strength, which was simply the log-likelihood PWM score at bound or unbound motifs. Also included in the set of motif-based features was the distance in base pairs to the closest TSS. Finally, a motif-based feature was constructed by calculating the average number of nearby motifs in a 200-bp window centered at every bound or unbound motifs; this analysis resulted in a total of 146 motif-based features.

2.3.3. Generation of sequence-based features

Of the 17 sequence-based features, 16 represented the normalized frequency of dinucleotides within a 200-bp window centered at bound or unbound motifs. For example, the feature measuring TA content would be calculated as the number of times the “TA” 2-mer was found within the 200-bp window divided by the number of k-mers of size 2 found in the window. Hence, this feature represents the enrichment of TA relative to all 2-mers. Also included was a sequence-based feature reflecting the overall content. Removing the reverse complement of a given dinucleotide (e.g., CG is the same as GC in the complementary strand) could further reduce the sequence features. Whether the reverse complement is redundant is based on whether strand-specific processes act at bound or unbound motifs. Since TF binding can be strand-specific, reverse complements were retained in the final set of features.

2.3.4. Feature selection

Feature selection can be described as finding the subset of features from the set of all possible combinations of features that can best distinguish classes of interest. In our case, the two classes of interest are bound and unbound motifs. Because the search space of all possible combinations of features grows exponentially with the number of features, it is rarely feasible to perform an exhaustive search. Instead, various heuristic search methods can be used to identify meaningful feature subsets. Here, we used three different feature selection algorithms to identify those features that are consistently selected by the different methods. We used two algorithms implemented in the open source software package ‘weka’ (Hall et al., 2009) and ‘galgo’, an R package (Treviso and Falciani, 2006).

Feature selection consists of two parts: the first are methods to score how well a feature subset predicts the correct class; the second are methods to search the space of all possible feature subsets and achieve convergence to an optimal feature subset.

The first 'weka'-based feature selection algorithm used was a correlation subset scoring approach paired with a best first search algorithm. Correlation subset scoring is based on the idea that a good subset of features contains those that are highly correlated with the class and yet do not correlate with each other (Hall, 1999). Feature subsets that have this characteristic are scored highly. This scoring function was paired with a best first search method, which searches the space of feature subsets using a greedy hill climbing approach augmented with backtracking.

The second 'weka'-based feature selection algorithm used was a consistency subset scoring approach paired with a linear forward selection search algorithm. Consistency subset scoring is based on the concept that the best features are those that are most consistent with a class. Thus, good features are consistently similar within a class but very different between classes (Liu and Setiono, 2000). This scoring function was paired with a linear forward search algorithm. Briefly, the search algorithm initially ranks all features individually using the consistency subset scoring method. Then the algorithm starts with an empty set of features and adds them one at a time based on ranking until performance can no longer be improved. The 'weka' version does allow for some backtracking and restarting of the search to help prevent quick convergence to small local optima. For more information, see Gutlein et al. (2009).

'Galgo' is a genetic algorithm-based feature selection approach. To score a feature subset, a nearest shrunken centroid classifier is built using only the feature subset. The score assigned to a feature subset represents how well the nearest shrunken centroid classifier performs in classifying held out test datasets of bound or unbound motifs (Trevino and Falciani, 2006).

All of the feature selection approaches used require a training dataset. Unfortunately, training datasets were frequently imbalanced, with far more examples of unbound motifs than bound motifs or vice versa. Many traditional machine learning-based approaches have lower accuracy when trained on imbalanced datasets (Japkowicz and Stephen, 2002). We performed repetitive random under-sampling to deal with the imbalanced dataset problem due to its simplicity and ability to improve performance (Van Hulse et al., 2007, 2009). Randomly removing examples from the majority class until balanced datasets are achieved is a common solution to the data imbalance problem. One drawback to this method is the possibility of discarding potentially useful information. Repetitive random under-sampling attempts to address this issue by combining the results from several rounds of random sampling. Studies have demonstrated that this method can potentially improve performance over simple under-sampling or not performing any sampling (Van Hulse et al., 2009; Xu-Ying et al., 2009).

We will describe our overall approach using the PHD1 dataset as an example. The PHD1 dataset is a highly imbalanced dataset consisting of 172 bound motifs and 1133 unbound motifs giving a total of 1305 motifs. For each motif, a vector of length 171 was constructed to produce a matrix with 1305 rows and 171 columns. Two-thirds of the rows representing bound and unbound motifs were randomly selected and set aside as a training dataset. This resulted in a training dataset with 114 rows representing bound motifs and 755 rows representing unbound motifs for a total dataset matrix of 869 rows with 171 columns. The remaining data was used as the test set. The training dataset then underwent random sampling without replacement selecting rows representing unbound motifs until a balanced dataset was achieved with 114 examples of bound motifs and 114 examples of randomly selected unbound motifs. This matrix was used as an input into the three feature selection methods, and the resulting feature subsets were stored. The process of randomly sampling from the training dataset to create a balanced dataset was repeated 10,000

times. The resulting 10,000 feature subsets were combined by counting the number of times each feature was observed. Features were then ranked based on the number of times each feature was selected. For example, if the "PWM_score" feature was included in 9000 out of the 10,000 feature subsets, it would rank higher than a feature selected in 1000 out of the 10,000 feature subsets. Each feature selection method produced a ranked list of features in this manner. Selecting features that were ranked in the top 10% by at least two of the three feature selection methods produced the final subset of features presented in Supplemental Table 1 and Supplemental Table 2. The above approach was used for each of the nine TFs that were analyzed by feature selection.

The features were globally ranked by pooling how often each feature was selected by each of the feature selection algorithms. Those features selected most often across all feature selection algorithms were presumed to be more important than features selected less often. Features listed in order of rank are provided in Supplemental Table 2.

Accuracy, sensitivity, and specificity were calculated by first producing a dataset using only those features selected using the feature selection approach. For example, the training dataset for PHD1 would consist of 869 rows and 16 columns with each column representing one of the 16 features listed in Supplemental Table 1. A balanced dataset was produced by random sampling and the resulting matrix was given as an input training dataset to build a random forest classifier (Breiman, 2001). The resulting classifier then works to correctly predict the class of motifs (bound or unbound) in the testing dataset. This procedure was repeated 10 times. The mean, accuracy, sensitivity, and specificity are presented in Supplemental Table 1. Hence, how well the features selected can discriminate between bound and unbound motifs was assessed on a testing dataset that was not used in feature selection.

3. Results

We obtained experimentally mapped binding sites in intergenic regions for 118 TFs (MacIsaac et al., 2006) in the yeast *S. cerevisiae* genome. We selected twelve TFs for further study since they have at least 100 experimentally mapped binding sites. A cutoff of 100 was used to ensure enough bound sites existed so useful statistics could be performed. For each of these, a set of motifs bound by the TF and a set of motifs likely unbound by the TF were obtained (see Materials and methods). DIG1's motif as described by MacIsaac et al. was present in only a minority of the experimentally-proven DIG1 binding sites ($n = 41$), suggesting the possibility of an error in the position weight matrix (PWM) used. As a result, DIG1 was not further analyzed in this study. For each of the motifs in the datasets, a 200-bp window centered on the motif and a vector containing 171 elements were calculated. Each element of the vector represented a measurement of a biological feature for that window. For example, vector element one is a score indicating the degree of nucleosome occupancy averaged over the 200-bp window. Feature selection was then applied to identify the subset of vector elements (hereby referred to as features) that were most informative in correctly predicting whether a motif is actually bound by the respective TF [for a review of the use of feature selection in bioinformatics, see Saeys et al., 2007]. We applied three different feature selection techniques (see Materials and methods), two of which were implemented in 'weka' (Hall et al., 2009) while the third is 'galgo' (Trevino and Falciani, 2006). Features selected by at least two of the three methods were examined in more detail (Supplemental Table 1). While feature selection can identify a set of promising candidates that differ between bound and unbound motifs, further analyses of the selected features are necessary to verify biologically significant differences. In general, there was good agreement between features selected by all three methods (Supplemental Fig. 1).

3.1. Correlation between motif strength and binding

In order to determine which biological features are associated with motifs bound by their TFs, it is necessary to compare the local environment of motifs bound by protein with motifs unbound by protein. Therefore our analysis required obtaining a dataset of motifs that are unlikely to be bound by a TF. We created this set for the 11 TFs examined from a published ChIP-chip dataset (Harbison et al., 2004), in which a p-value was calculated representing the degree of evidence regarding binding to each intergenic region in the yeast genome (in general the lower the p-value the stronger the ChIP-chip evidence the given TF binds somewhere in the intergenic region). Our unbound motif dataset consisted of motifs that occurred in intergenic regions with a p-value greater than 0.5 in all experimental conditions studied by Harbison et al. Using these criteria, we discovered that two out of the 11 TFs, CBF1 and ABF1, exhibited too few non-bound motif matches (8 and 38, respectively); therefore, these TFs were eliminated from further feature selection analysis.

To further explore the relationship between the presence of a motif match and the p-value for binding as measured by Harbison et al., we plotted the average p-values for intergenic regions that contain strong motif matches (*i.e.*, matches > 80% of the maximum possible PWM log likelihood score; Fig. 1, panel a). Although there was low information content for the majority of the motifs, the presence of a strong motif was a surprisingly good predictor of binding for many of the TFs (Fig. 1, panel a). ABF1 and CBF1 had the two lowest average p-values for binding 0.022 and 0.044, respectively.

It is reasonable to expect a connection between motifs with high information content and a higher probability of binding to a strong motif. Indeed, a positive correlation between information content and p-value for binding to strong motifs ($r = -0.67$, $p\text{-value} = 0.02$) was observed (Fig. 1, panel b). Next, we plotted the average p-value for binding at different motif strengths for ABF1 and SUT1 the two extreme cases (Fig. 1, panels c and d). ABF1 showed an almost perfect correlation between motif strength and p-value ($r = -0.98$, $p\text{-value} = 0.00009$). In contrast, a positive correlation was found for SUT1 ($r = 0.66$, $p\text{-value} = 0.1055$); however, this correlation was not statistically significant ($\alpha = 0.05$).

3.2. Comparison of sequence-based features surrounding bound and unbound motifs

Some of the features assessed by the feature selection algorithms were sequence-based (*e.g.*, dinucleotide content, see Supplemental Table 1). To explore this further, we plotted the percentage of dinucleotides surrounding bound and unbound motifs (Fig. 2 and Supplemental Fig. 3). In this analysis, we masked the actual motif and 15 bp flanking both sides. The percentage of the dinucleotide TA present near bound and unbound motifs was calculated as the ratio of TAs present in a given sequence (Fig. 2). Out of the 16 dinucleotides examined, TA was selected by our feature selection approach for all nine TFs as an important feature that discriminates between bound and unbound motifs (Supplemental Table 1, Supplemental Table 2). The peak observed in the control dataset is due to the fact that

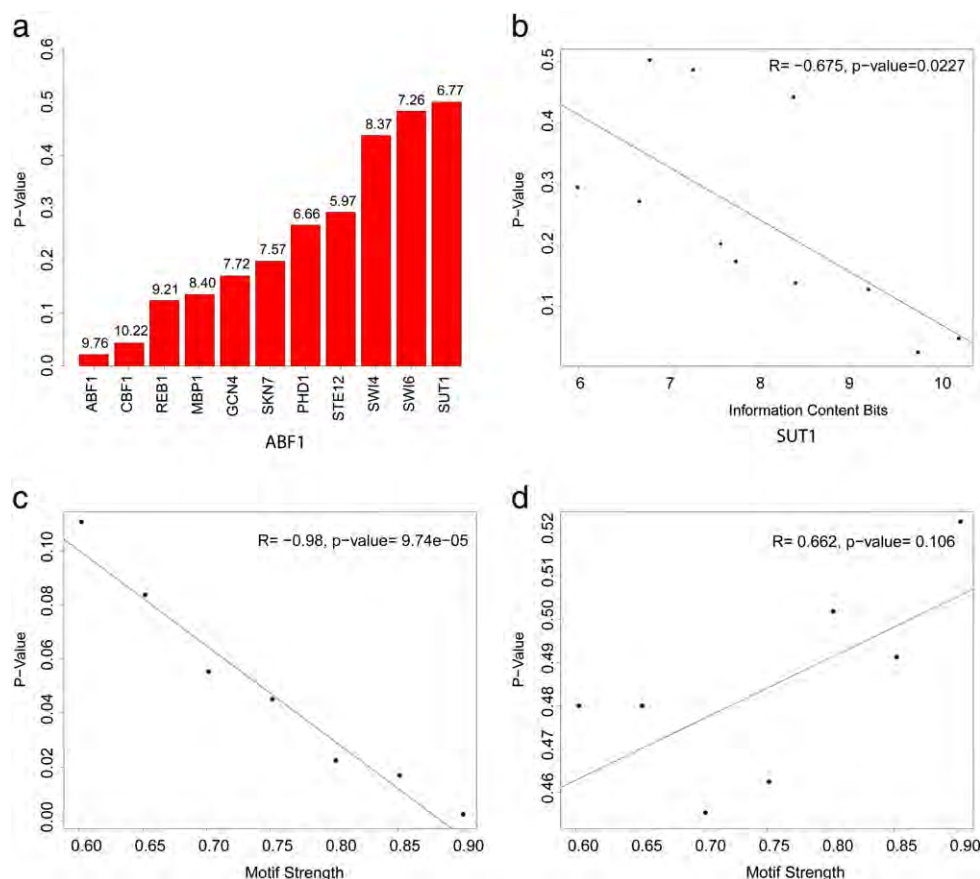


Fig. 1. Correlation between motif strength and p-value of binding. (a) Plotted is the mean p-value of binding for intergenic regions whose average motif strength was > 80% of the maximum possible log-likelihood score. The p-value of binding was obtained from Harbison et al. (2004). The number above each bar is the information content for the given motif in bits. The smaller the information content, the more likely that motif is to occur by random chance in a sequence. (b) For every motif, the average p-value of binding in intergenic regions containing high scoring motifs was calculated as described above (y-axis). The x-axis is the information content of the motifs in bits. (c and d) Plots of the p-value of binding versus motif strength for (c) ABF1 and (d) SUT1. The x-axis denotes the motif strength of a given TF as a percentage of the maximum possible PWM log-likelihood score. Higher motif strength correlates with closer proximity to the consensus sequence. The average p-value of binding for the collected intergenic regions that met the given motif strength threshold was calculated (y-axis). ABF1 and SUT1 were plotted because they represent the two extremes.

intergenic regions are TA-rich compared to coding regions (Fig. 2, green line).

In general, the sequence surrounding bound motifs was depleted of TA dinucleotides compared to unbound motifs (Fig. 2) with six (SWI4, PHD1, SKN7, SUT1, STE12, and SWI6) out of the nine TFs clearly showing this pattern. Two TFs, MBP1 and GCN4, did not exhibit strong differences in the percentage of TA between bound and unbound motifs. REB1 showed slightly higher levels of the TA dinucleotide around bound sites compared to unbound sites. SWI6 may not bind DNA directly but instead be recruited to the genes it regulates by other TFs. It is known that SWI4 and MBP1 can bind to DNA as a complex with SWI6 (Andrews and Moore, 1992; Leem et al., 1998). Given the indirect binding of SWI6 to DNA, it is likely the motif identified for SWI6 is a combination of the motifs recognized by the proteins that recruit SWI6 to DNA. Indeed, the core SWI6 motif CGCG is found in both the SWI4 motif and the MBP1 motif (Supplemental Table 3). Furthermore, the local TA dinucleotide content surrounding the bound SWI6 motif closely resembles that of bound SWI4 motifs (Fig. 2). Many genes have a tendency to be regulated by multiple TFs. Hence it is possible that the differences in sequence composition when comparing bound to unbound motifs are due to bound motifs having multiple other motifs nearby which may affect local sequence composition. To control for this we obtained all motifs with some evidence of being bound by protein (MacIsaac et al., 2006) and masked these motifs.

Additionally, a number of TFs show the same general trend with regard to differences in sequence composition when comparing bound to unbound motifs. The example given above is the six TFs that all show the same pattern of depleted TA dinucleotides around bound motifs compared to unbound. Since it is unlikely these six TFs all share the same regulatory partners, it is unlikely that the

same pattern of depleted TA dinucleotides is due to the potential confounding effect of having multiple other bound TF motifs nearby. The same motifs will likely not be present around all six TFs since they do not share the same regulatory partners (see Fig. 3).

Several studies have shown TFs in yeast exhibiting distinct positional preferences relative to the transcription start site (TSS) (Hansen et al., 2010; Harbison et al., 2004; Lin et al., 2010). Thus, it is possible that the differences in dinucleotide content are due to bound motifs predominantly occurring –100 to –500 bp upstream of the TSS (Harbison et al., 2004). To investigate this, we extracted the noncoding sequence –100 to –500 bp upstream of all yeast TSSs and calculated the TA dinucleotide content for these regions. The percentage of TA was slightly lower in sequences –100 to –500 bp upstream of the TSS than in intergenic regions as a whole (0.091 compared to 0.099). However, this phenomenon was insufficient to explain the pronounced depletion of TA around bound motifs found for many of the TFs (Fig. 2). For example, the average percentage of TA within a 200-bp window centered at bound SUT1 motifs was 0.054. Hence, reduced TA content is not universal throughout promoter regions, but is instead generally found in sequences surrounding motifs bound by TFs. Additionally TFs, in general share, similar location binding preferences, but do not always share the same pattern of dinucleotide frequency around bound motifs. For example PHD1 prefers to bind on average ~340 bp upstream of the TSS while SWI4 prefers to bind on average ~380 bps upstream of a TSS. PHD1's bound motifs in general are not embedded in GG rich sequence, while SWI4s motifs are (Supplemental Fig. 2).

While for TA the trend is for bound motifs to be embedded in TA depleted sequence compared to unbound motifs, this is not the case for other dinucleotides. For example, the GG dinucleotide shows a tendency to be enriched around bound motifs relative to unbound motifs (Supplemental Fig. 2). In general, the dinucleotide content of

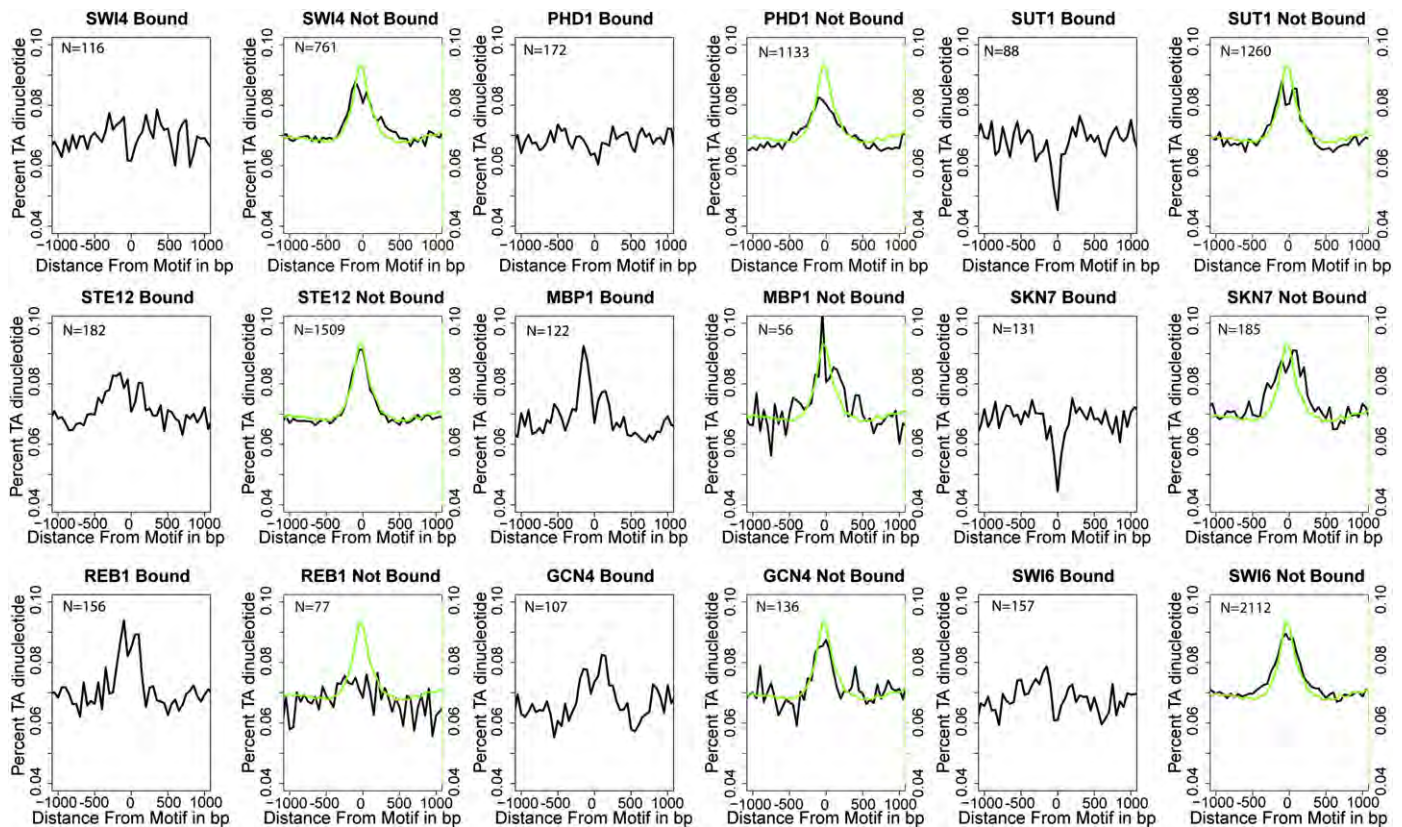


Fig. 2. TA dinucleotide content around bound or unbound motifs. Motifs classified as bound or unbound were aligned. The TA dinucleotide content was bound in 50-bp windows moving upstream and downstream from the motif. Zero on the x-axis represents the center of the aligned motif. Black: The average percentage of TA, which is defined as the fraction of dinucleotides that are TA within each 50 bp window. Green: The background TA content calculated by randomly selecting locations in intergenic regions and repeating the procedure as described.

unbound motifs was similar to the overall background intergenic content, while the dinucleotide content around bound motifs was either enriched or depleted relative to background (with the exception of REB1). Given the apparent strong dependence of ABF1 on its motif, we examined the dinucleotide content surrounding its bound motifs relative to background (Supplemental Fig. 2). Contrary to the trend observed for many of the other nine TFs, the dinucleotide content surrounding bound ABF1-specific motifs does not show strong deviations from the background dinucleotide content.

3.3. Comparison of motif-based features surrounding bound and unbound motifs

For all nine TFs, the feature selection algorithms selected the distance from the motif to the nearest TSS as a significant discriminator between bound and unbound motifs. Many of the TFs exhibited a striking difference between bound and unbound motifs. For instance, the median distance to the nearest TSS for PHD1 was -430 and -155 for bound and unbound motifs, respectively. This result was expected given the strong positional preference relative to the TSS seen for many yeast TFs (Hansen et al., 2010; Harbison et al., 2004; Kim et al., 2008; Lin et al., 2010; Tharakaraman et al., 2005).

Included in the set of motif-based features were a number of characteristics designed to take advantage of the fact that TFs have a tendency to bind in clusters (Frith et al., 2003; Ptashne, 1988). To construct these features, we obtained a library of PWMs representing 143 TFs (Badis et al., 2008; Zhu et al., 2009). For each bound and unbound motif, we counted the number of motif matches within 100 bp for each of the TFs in our PWM library. By comparing the number of motif matches near bound and unbound sites, we identified motifs that are commonly found near bound and unbound motifs (Fig. 3). Interestingly, three TFs (MBP1, STE12, and SWI4) bound motifs had a

tendency to have repeated copies of their motifs surrounding their binding sites. Such homotypic clusters of the same motif have been observed in other organisms including vertebrates and invertebrates (Gotea et al., 2010; Lifanov et al., 2003). As our results and others (Harbison et al., 2004) have shown, homotypic clustering is also present in yeast, suggesting an evolutionarily conserved regulatory mechanism.

Regulation of transcription initiation is facilitated by the binding of multiple TFs to the promoter region of a gene. Indeed many of the TFs whose motifs are enriched at bound sites relative to unbound sites showed signs of cooperative binding. For example, in budding yeast numerous genes are induced early in the cell cycle with SWI4 and MBP1 as the predominant regulators of these genes (Koch et al., 1993; Sidorova and Breeden, 1993). In some cases these genes cooperate in regulating the same gene (Bean et al., 2005). Unsurprisingly we observed enrichment of MBP1 motifs surrounding SWI4-bound motifs compared to unbound motifs. SWI4 motifs were also enriched around MBP1-bound motifs compared to unbound motifs; however, this enrichment (q -value = 0.07) did not meet our q -value cutoff of 0.05 (Fig. 3). In addition, enrichment of the TEC1 motif near bound STE12 motifs exhibited a similar pattern. STE12 is necessary for the proper regulation of mating, haploid invasion, and pseudohyphal development (Herskowitz, 1995). STE12 binds with TEC1 cooperatively to achieve developmental specificity (Madhani and Fink, 1997), which is consistent with our observation that the TEC1 motif is enriched around STE12-bound motifs compared to unbound ones.

Since TFs have a tendency to bind cooperatively, a greater number of motifs can be found enriched around bound motifs compared to unbound ones. Therefore, the feature measuring the average number of nearby motifs was selected for five out of the nine TFs (MBP1, SKN7, STE12, SWI4, and SWI6), further indicating the tendency for the enrichment of multiple motifs surrounding TF binding sites.

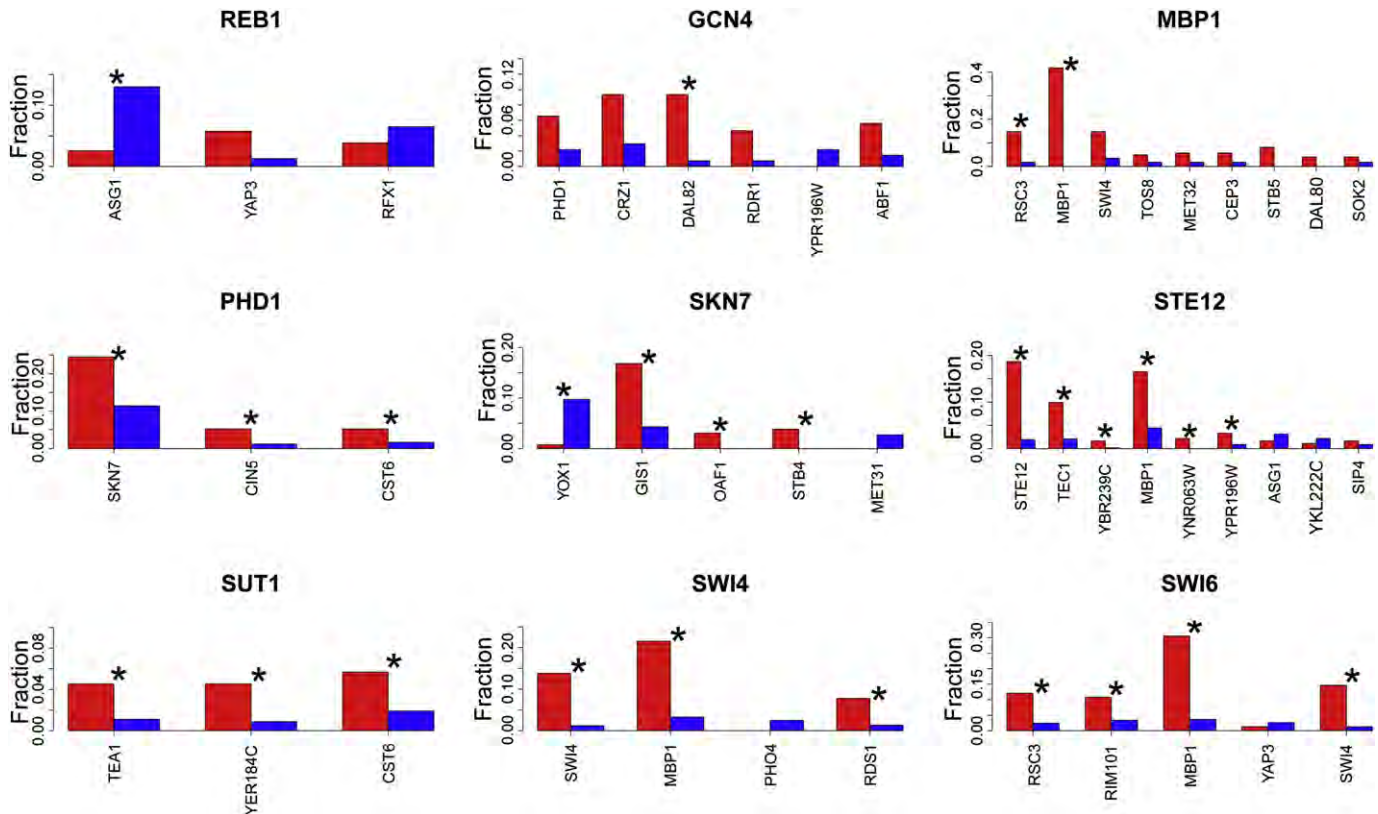


Fig. 3. Motifs enriched near bound or unbound motifs. The fraction of bound (red) or unbound (blue) motifs that exhibit at least one of the labeled motifs within 100 bp is plotted for the nine TFs shown. p -Values were calculated using the z -test for two proportions, and corrected for multiple testing using Benjamini, Hochberg, and Yekutieli correction (Benjamini and Yekutieli, 2001). Comparisons with a q -value < 0.05 are marked with an asterisk.

However, there were also instances of enrichment of motifs around unbound motifs compared to bound motifs. For example, four of the nine TFs (GCN4, PHD1, SKN7, and SWI4) exhibited statistically significant enrichment of the PHO2 motif around unbound motifs compared to bound ones. This unique enrichment may occur because unbound motifs are generally located within TA-rich regions of the genome (Fig. 2). Because the PHO2 motif is TA/AT-rich and information poor, it is not surprising that this motif is widespread across yeast intergenic regions (~32,000 PHO2 motifs in intergenic regions $N=32,562$). This widespread occurrence may explain why the PHO2 motif, as well as those of SIG1 and GLN3, is enriched near unbound motifs. To control for the tendency of information poor motifs to be strongly influenced by local sequence context we filtered motifs on the basis of information content. And only counted motifs near bound and not bound motifs with at least 8 bits of information. This filtering removed the PHO2, SIG1 and GLN3 motifs from consideration (Fig. 3).

However, the above explanation does not account for all cases of motif enrichment around unbound motifs. For instance, despite lacking in AT/TA dinucleotides, the ASG1 motif is enriched around unbound REB1 and STE12 motifs. Given the enrichment of the ASG1 motif around unbound motifs it is possible that ASG1 may be acting as a repressor.

The average motif strength for bound and unbound motifs (*i.e.*, feature “PWM_score,” Supplemental Table 1 and Supplemental Table 2) was selected for all TFs but SKN7 and SUT1. On average, bound motifs were stronger than unbound motifs.

3.4. Comparison of nucleosome-based features surrounding bound and unbound motifs

REB1 possesses nucleosome-modifying properties and functions to create regions of open chromatin (Chasman et al., 1990). Hence, nucleosome occupancy was selected as an important feature to discriminate between REB1-bound and unbound motifs, with bound motifs located in nucleosome-depleted regions (data not shown). Interestingly, although nucleosome occupancy was selected as an important feature for SKN7, bound sites had a higher nucleosome occupancy score than unbound sites (data not shown). This result is contrary to the overall trend of bound sites occurring predominantly in nucleosome-depleted regions (Kaplan et al., 2009). Nevertheless, a recent study mapping nucleosome occupancy suggested that the presence of SKN7 leads to higher nucleosome occupancy at its binding site (Kaplan et al., 2009).

Several histone post-translational modifications have been associated with either bound or unbound motifs. Often, the level of histone modification is associated closely with gene activity (Pokholok et al., 2005). Hence a correlation between active binding sites and histone modification levels is expected. Surprisingly, we also observed an enrichment of active histone marks reported to be associated with active genes, namely H3K4me3, H3K9ac, and H3K14ac (Pokholok et al., 2005), around unbound motifs. Our data demonstrate that these marks are enriched around unbound motifs for a number of TFs (PHD1, SKN7, SUT1, and SWI4) (Fig. 4).

These histone marks are present at the highest levels near the TSS (Pokholok et al., 2005). Consistent with this, higher levels of histone modification can be found around motifs that are closest to the TSS. TFs have a tendency to avoid binding 0 to approximately 100 bp upstream of the TSS (Harbison et al., 2004; Lin et al., 2010). Thus, enrichment of active histone marks around unbound motifs may occur because a larger fraction of these motifs is located within 0 to 100 bp of the TSS. Indeed, our data supports this hypothesis. The percentage of bound sites within 100 bp of the TSS for PHD1, SKN7, SUT1, and SWI4 was 5.24%, 4.58%, 1.14%, and 1.72%, respectively. Meanwhile, the percentage of unbound motifs within 100 bps upstream of the TSS for PHD1, SKN7, SUT1, and SWI4 was 21.27%, 13.51%,

12.78%, and 19.06%, respectively. Because the region 0 to approximately 140 bp upstream of the TSS is free of nucleosomes (Kaplan et al., 2009; Lee et al., 2007; Shivaswamy et al., 2008), motifs located within this region are most likely in an open chromatin configuration and accessible for TF binding, which raises the question what mechanism is repressing binding at these motifs.

4. Discussion

Given the prominent role TFs play in gene regulation throughout the genome, mapping TF binding is very important to gaining a thorough understanding of transcription. In recent years, ChIP-chip and ChIP-seq have become widely used experimental tools in identifying binding locations for TFs. Unfortunately, even with these techniques, mapping the binding sites for large numbers of TFs is still a substantial undertaking. Computational prediction of TF binding sites has the potential to provide high-quality predictions of TF binding with precision and low cost. Indeed this has been an active area of research for computational biologists (Elemento and Tavazoie, 2005; Ernst et al., 2010; Pique-Regi et al., 2011; Xie et al., 2009). Our primary goal in this analysis is not prediction of TF binding site but identifying differences in local genomic context comparing bound motifs to unbound motifs.

We examined a subset of yeast TFs so our results cannot be generalized across all TFs. While it is true the total number of TFs studied was low in comparison to the total number of TFs in yeast. There was high diversity in the DNA binding domains; the nine different TFs represent six different DNA binding domain families (Supplemental Table 3). There were several broad trends that were universal or mostly universal across all TFs studied. For example, every TF examined in this study showed differences in local sequence composition around bound motifs compared to unbound motifs. With the sequence composition surrounding unbound motifs in general corresponding to the background sequence composition.

Experimental data suggests local sequence content may be important in transcription factor binding site functioning (Meierhans et al., 1997; Ponomarenko et al., 1999; Starr et al., 1995). Our results are consistent with these findings. However it is not apparent what role sequence context plays in transcription factor binding. In some cases it is clear that sequence context plays a direct role in stabilizing binding (Meierhans et al., 1997; Starr et al., 1995). It is also possible that sequence context is important indirectly through mediating nucleosome binding. Indeed nucleosome occupancy around TF binding sites is depleted of nucleosomes *in vitro* strongly suggesting sequence context plays a role in excluding nucleosomes (Kaplan et al., 2009). Interestingly, sequences containing 9–11 bp periodic TA dinucleotides have recently been shown *in vitro* to have strong nucleosome forming potential (Takasuka and Stein, 2010).

The TA dinucleotide was identified by our approach as being important for all 9 TFs in distinguishing between bound and unbound motifs. In general sequences around bound motifs were depleted of the TA dinucleotide compared to unbound motifs (Fig. 2); this effect is stronger for some TFs than others. An exception to this general trend is the REB1 motif which showed increased TA frequency around bound motifs compared to unbound motifs. The REB1 protein has chromatin modifying properties with the ability to form nucleosome free regions (Chasman et al., 1990). Indeed if TA dinucleotides in certain sequence contexts increase nucleosome formation, the depletion of TA dinucleotides around bound motifs we observe would presumably discourage nucleosome formation.

It is however unlikely that the differences in sequence context comparing bound to unbound motifs are entirely explained by nucleosome sequence preferences. The TFs examined do not always exhibit consistent sequence preferences. For example SWI4, SUT1 and SKN7 all have enriched GG dinucleotide content around bound motifs while STE12 and PHD1 do not (Supplemental Fig. 3). Nucleosome

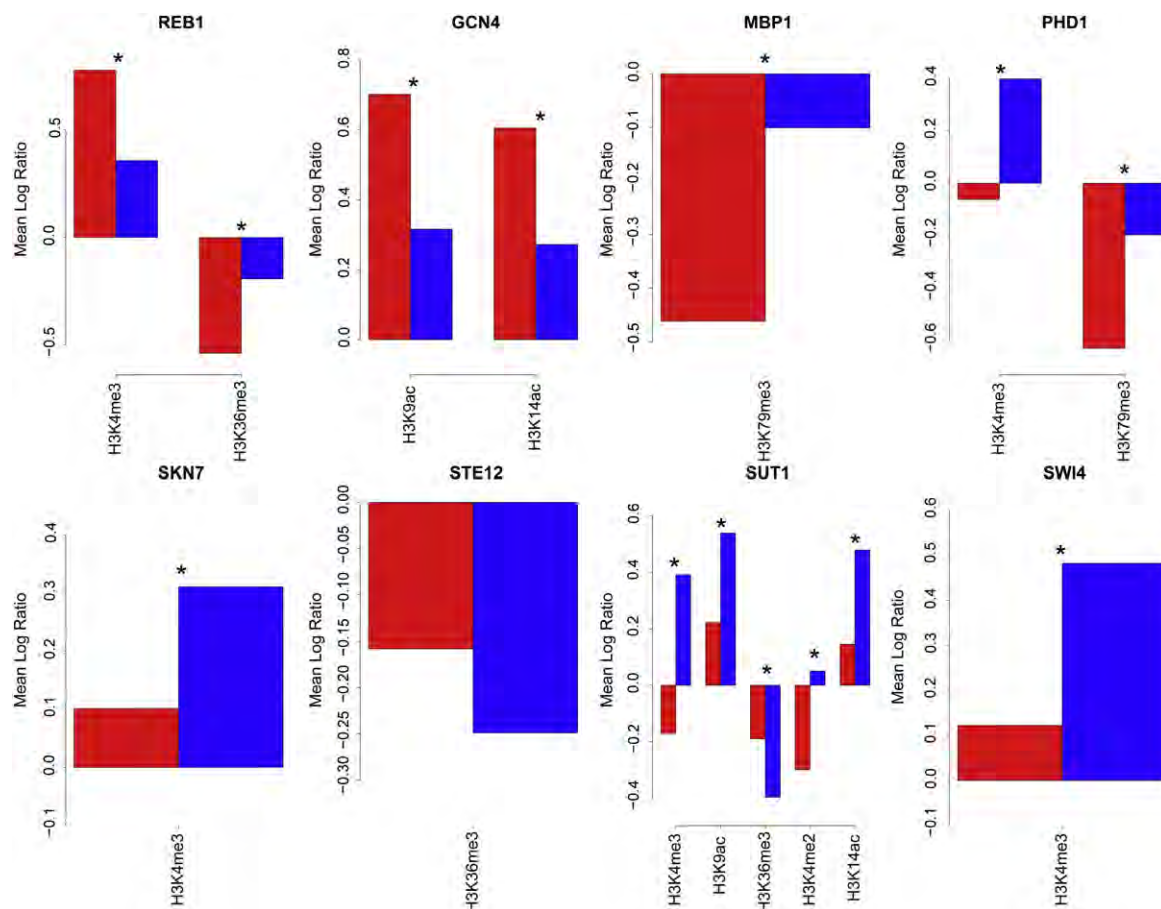


Fig. 4. Histone modification-based features. Histone modification-based features are plotted for the eight TFs for which a histone modification feature was selected as important. Red bars represent the average log ratio of the given histone modification within a 200-bp window centered at bound sites. Blue bars represent the average value of the given nucleosome-based feature within a 200-bp window centered at unbound sites. p-Values were calculated using the Wilcoxon rank sum test, and corrected for multiple testing using the Benjamini, Hochberg, and Yekutieli correction (q-values) (Benjamini and Yekutieli, 2001). Comparisons with a q-value < 0.05 are marked with an asterisk.

sequence preferences should theoretically be consistent within a cell, hence if the differences in sequence composition observed are entirely due to nucleosome sequence preference this preference would be expected to be consistent for all TFs.

Another possible explanation for the differences in sequence composition comparing bound to unbound motifs is direct stabilization of binding. The recognition of binding locations by DNA binding proteins is dependent on two different approaches: first nucleotide sequence specific formation of hydrogen bonds and second nonbase pair specific interactions between the protein body and DNA (Rohs et al., 2009). It has recently been shown that the binding of arginine residues to narrow minor grooves is a common mechanism assisting in protein–DNA recognition (Rohs et al., 2009). Differences in sequence composition around an embedded motif could either enhance or inhibit such interactions by affecting the DNA shape or width of the major/minor groove. It would be of interest to measure the binding affinity of the same motif embedded in different sequence contexts.

It is also possible that differences in sequence context between bound and unbound motifs are reflective of differences in sequence composition between regions of regulatory sequence and non-regulatory sequence. While this is a possibility we observed there is little difference in TA dinucleotide composition in promoter sequence compared to background yeast intergenic regions. This is not surprising since in yeast, intergenic regions are compact with promoter sequence being a large fraction of intergenic sequence. This suggests that the differences in dinucleotide frequency comparing bound to unbound motifs are not due to a general trend observed in regulatory sequence.

Understanding gene regulation is a fundamental question in molecular biology. Many genes are regulated by TFs recognizing and binding to short DNA sequence motifs. In most cases, only subsets of the genomic regions that match TF binding sites are actually bound by the TF *in vivo*. Thus, it is critical to understand the difference between motifs that are bound and unbound by a given TF. Here, we begin to investigate this question by performing a systematic genome-wide comparison of motifs that are bound *in vivo* compared to motifs that are unbound. To our knowledge, this is the first such study. Further work could extend the set of biological features being examined. Our analysis cannot answer whether any of the differences we identify are a causal component of TF binding specificity.

For ABF1 and CBF1, our results suggest that the presence of a strong motif is a good predictor of binding in intergenic regions. Both proteins have chromatin-modifying properties (Yarragudi et al., 2004; Kent et al., 2004). In agreement with our results, genome localization studies indicate that the majority of CBF1 motifs in intergenic regions are most likely bound by the TF (Kent et al., 2004; Lee et al., 2002).

Our results suggest a range of strategies is employed in determining DNA binding specificity. For some TFs (e.g. ABF1 and CBF1 (Fig. 1)) the information contained in their motif is apparently sufficient to mostly determine specificity. Little additional information from genomic context is needed. Every place a strong copy of their motif is found it may be likely the protein will bind. We reported previously that the ABF1 motif is strongly biased to occur predominantly in potential regulatory regions. We also showed that the ABF1 motif exhibits a strong positional preference relative to the TSS (Hansen et al., 2010).

For other TFs whose motifs are information poor and found in high abundance throughout the genome the information contained in the motif is not sufficient to determine specificity and input from the local genomic environment may play a dominant role in determining specificity. The majority of TFs may fall somewhere between these two scenarios depending to a greater or lesser extent on genomic context to determine specificity.

It is also likely there is interplay between these two approaches. ABF1 is an abundant general regulatory factor essential to cell growth (Halfter et al., 1989). This factor acts in part by creating a bubble of open chromatin (Yarragudi et al., 2004). In many cases, ABF1 alone is insufficient to activate robust transcription and requires the cooperation of other regulatory factors (Goncalves et al., 1995). A recent study indicates that ABF1 may play an important role in determining chromatin structure throughout the genome, with weaker motifs showing evidence of ABF1 binding and chromatin remodeling (Ganapathi et al., 2011). Genome-wide interaction studies have identified ABF1 to be a network “hub,” suggesting that it plays a central role in gene regulation (Zhang et al., 2006).

Given these results, ABF1 may act in part as an important pioneer TF that binds chromatin and acts to create regions of open chromatin that allows other factors to bind similar to pioneer factors in higher organisms (Zaret et al., 2008). ABF1 could be acting to create a local genomic environment conducive for other TFs to bind, while ABF1's binding specificity is dependent mostly on the presence of its motif.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine and National Center for Biotechnology Information.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.gene.2012.06.005>.

References

- Andrews, B.J., Moore, L.A., 1992. Interaction of the yeast Swi4 and Swi6 cell cycle regulatory proteins *in vitro*. *Proc. Natl. Acad. Sci. U. S. A.* 89, 11852–11856.
- Badis, G., et al., 2008. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell* 32, 878–887.
- Bauer, A.L., Hlavacek, W.S., Unkefer, P.J., Mu, F., 2010. Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites. *PLoS Comput. Biol.* 6, e1001007.
- Bean, J.M., Siggia, E.D., Cross, F.R., 2005. High functional overlap between MluI cell-cycle box binding factor and Swi4/6 cell-cycle box binding factor in the G1/S transcriptional program in *Saccharomyces cerevisiae*. *Genetics* 171, 49–61.
- Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188.
- Blanchette, M., Tompa, M., 2003. FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.* 31, 3840–3842.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32.
- Chasman, D.I., Lue, N.F., Buchman, A.R., LaPointe, J.W., Lorch, Y., Kornberg, R.D., 1990. A yeast protein that influences the chromatin structure of UASG and functions as a powerful auxiliary gene activator. *Genes Dev.* 4, 503–514.
- D'Haeseleer, P., 2006. What are DNA sequence motifs? *Nat. Biotechnol.* 24, 423–425.
- Elemento, O., Tavazoie, S., 2005. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.* 6, R18.
- Ernst, J., Plasterer, H.L., Simon, I., Bar-Joseph, Z., 2010. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.* 20, 526–536.
- Fickett, J.W., 1996. Quantitative discrimination of MEF2 sites. *Mol. Cell. Biol.* 16, 437–441.
- Frith, M.C., Li, M.C., Weng, Z., 2003. Cluster-buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* 31, 3666–3668.
- Ganapathi, M., et al., 2011. Extensive role of the general regulatory factors, Abf1 and Rap1, in determining genome-wide chromatin structure in budding yeast. *Nucleic Acids Res.* 39, 2032–2044.
- Goncalves, P.M., et al., 1995. Transcription activation of yeast ribosomal protein genes requires additional elements apart from binding sites for Abf1p or Rap1p. *Nucleic Acids Res.* 23, 1475–1480.
- Gotea, V., Visel, A., Westlund, J.M., Nobrega, M.A., Pennacchio, L.A., Ovcharenko, I., 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.* 20, 565–577.
- Gutlein, M., Frank, E., Hall, M., Karwath, A., 2009. Large-scale attribute selection using wrappers. *Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on*, pp. 332–339.
- Halfter, H., Kavety, B., Vandekerckhove, J., Kiefer, F., Gallwitz, D., 1989. Sequence, expression and mutational analysis of BAF1, a transcriptional activator and ARS1-binding protein of the yeast *Saccharomyces cerevisiae*. *EMBO J.* 8, 4265–4272.
- Hall, M., 1999. Correlation-based Feature Subset Selection for Machine Learning. Department of Computer Science, University of Waikato.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* 11, 10–18.
- Hansen, L., Marino-Ramirez, L., Landsman, D., 2010. Many sequence-specific chromatin modifying protein-binding motifs show strong positional preferences for potential regulatory regions in the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.* 38, 1772–1779.
- Harbison, C.T., et al., 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104.
- Herskowitz, I., 1995. MAP kinase pathways in yeast: for mating and more. *Cell* 80, 187–197.
- Holloway, D.T., Kon, M., DeLisi, C., 2005. Integrating genomic data to predict transcription factor binding. *Genome Inform.* 16, 83–94.
- Holloway, D.T., Kon, M., DeLisi, C., 2007. Machine learning for regulatory analysis and transcription factor target prediction in yeast. *Syst. Synth. Biol.* 1, 25–46.
- Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. *Intell. Data Anal.* 6, 429–449.
- Kaplan, N., et al., 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458, 362–366.
- Kent, N.A., Eibert, S.M., Mellor, J., 2004. Cbf1p is required for chromatin remodeling at promoter-proximal CACGTG motifs in yeast. *J. Biol. Chem.* 279, 27116–27123.
- Kim, N.K., Tharakaraman, K., Marino-Ramirez, L., Spouge, J.L., 2008. Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. *BMC Bioinformatics* 9, 262.
- Koch, C., Moll, T., Neuberg, M., Ahorn, H., Nasmyth, K., 1993. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science* 261, 1551–1557.
- Lee, T.I., et al., 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804.
- Lee, W., et al., 2007. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* 39, 1235–1244.
- Leem, S.H., Chung, C.N., Sunwoo, Y., Araki, H., 1998. Meiotic role of SWI6 in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 26, 3154–3158.
- Lifanov, A.P., Makeev, V.J., Nazina, A.G., Papatsenko, D.A., 2003. Homotypic regulatory clusters in *Drosophila*. *Genome Res.* 13, 579–588.
- Lin, Z., Wu, W.S., Liang, H., Woo, Y., Li, W.H., 2010. The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcriptional regulation. *BMC Genomics* 11, 581.
- Liu, H., Setiono, R., 2000. A Probabilistic Approach to Feature Selection — A Filter Solution, pp. 319–327.
- MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., Stormo, G.D., Fraenkel, E., 2006. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* 7, 113.
- Madhani, H.D., Fink, G.R., 1997. Combinatorial control required for the specificity of yeast MAPK signaling. *Science* 275, 1314–1317.
- Meierhans, D., Sieber, M., Allemann, R.K., 1997. High affinity binding of MEF-2C correlates with DNA bending. *Nucleic Acids Res.* 25, 4537–4544.
- Peng, S., Alekseyenko, A.A., Larschan, E., Kuroda, M.I., Park, P.J., 2007. Normalization and experimental design for ChIP-chip data. *BMC Bioinformatics* 8, 219.
- Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., Pritchard, J.K., 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* 21, 447–455.
- Pokholok, D.K., et al., 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 122, 517–527.
- Ponomarenko, J.V., Ponomarenko, M.P., Frolov, A.S., Vorobyev, D.G., Overton, G.C., Kolchanov, N.A., 1999. Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics* 15, 654–668.
- Ptashne, M., 1988. How eukaryotic transcriptional activators work. *Nature* 335, 683–689.
- Rhee, H.S., Pugh, B.F., 2011. Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. *Cell* 147, 1408–1419.
- Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., Honig, B., 2009. The role of DNA shape in protein–DNA recognition. *Nature* 461, 1248–1253.
- Saeyns, Y., Inza, I., Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517.
- Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M., Iyer, V.R., 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* 6, e65.
- Sidorova, J., Breeden, L., 1993. Analysis of the SWI4/SWI6 protein complex, which directs G1/S-specific transcription in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 13, 1069–1077.
- Starr, D.B., Hoopes, B.C., Hawley, D.K., 1995. DNA bending is an important component of site-specific recognition by the TATA binding protein. *J. Mol. Biol.* 250, 434–446.
- Takasuka, T.E., Stein, A., 2010. Direct measurements of the nucleosome-forming preferences of periodic DNA motifs challenge established models. *Nucleic Acids Res.* 38, 5672–5680.

- Tharakaraman, K., Marino-Ramirez, L., Sheetlin, S., Landsman, D., Spouge, J.L., 2005. Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics* 21 (Suppl. 1), i440–i448.
- Tharakaraman, K., Bodenreider, O., Landsman, D., Spouge, J.L., Marino-Ramirez, L., 2008. The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. *Nucleic Acids Res.* 36, 2777–2786.
- Trevino, V., Falciani, F., 2006. GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics* 22, 1154–1156.
- Van Hulse, J., Khoshgoftaar, T., Napolitano, A., 2007. Experimental perspectives on learning from imbalanced data. *ICML '07: Proceedings of the 24th International Conference on Machine Learning*. ACM, Corvallis, Oregon, pp. 935–942.
- Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A., 2009. An empirical comparison of repetitive undersampling techniques. *Information Reuse & Integration*, 2009. IRI '09. IEEE International Conference on, pp. 29–34.
- Wasserman, W.W., Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 5, 276–287.
- Xie, X., et al., 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434, 338–345.
- Xie, X., Rigor, P., Baldi, P., 2009. MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics* 25, 167–174.
- Xu-Ying, L., Jianxin, W., Zhi-Hua, Z., 2009. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* 39, 539–550.
- Yarragudi, A., Miyake, T., Li, R., Morse, R.H., 2004. Comparison of ABF1 and RAP1 in chromatin opening and transactivator potentiation in the budding yeast *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 24, 9152–9164.
- Zaret, K.S., Watts, J., Xu, J., Wandzioch, E., Smale, S.T., Sekiya, T., 2008. Pioneer factors, genetic competence, and inductive signaling: programming liver and pancreas progenitors from the endoderm. *Cold Spring Harb. Symp. Quant. Biol.* 73, 119–126.
- Zhang, Z., et al., 2006. Dynamic changes in subgraph preference profiles of crucial transcription factors. *PLoS Comput. Biol.* 2, e47.
- Zhu, C., et al., 2009. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.* 19, 556–566.

Chapter 17

The Practical Evaluation of DNA Barcode Efficacy*

John L. Spouge and Leonardo Mariño-Ramírez

Abstract

This chapter describes a workflow for measuring the efficacy of a barcode in identifying species. First, assemble individual sequence databases corresponding to each barcode marker. A controlled collection of taxonomic data is preferable to GenBank data, because GenBank data can be problematic, particularly when comparing barcodes based on more than one marker. To ensure proper controls when evaluating species identification, specimens not having a sequence in every marker database should be discarded. Second, select a computer algorithm for assigning species to barcode sequences. No algorithm has yet improved notably on assigning a specimen to the species of its nearest neighbor within a barcode database. Because global sequence alignments (e.g., with the Needleman–Wunsch algorithm, or some related algorithm) examine entire barcode sequences, they generally produce better species assignments than local sequence alignments (e.g., with BLAST). No neighboring method (e.g., global sequence similarity, global sequence distance, or evolutionary distance based on a global alignment) has yet shown a notable superiority in identifying species. Finally, “the probability of correct identification” (PCI) provides an appropriate measurement of barcode efficacy. The overall PCI for a data set is the average of the species PCIs, taken over all species in the data set. This chapter states explicitly how to calculate PCI, how to estimate its statistical sampling error, and how to use data on PCR failure to set limits on how much improvements in PCR technology can improve species identification.

Key words: Barcode efficacy in species identification, Probability of correct identification, DNA barcode

1. Introduction

Species are becoming extinct, making conservation of biodiversity a major challenge. The first step to preserving biodiversity is assessment, but there are not enough taxonomists to catalog species

*For software relevant to this chapter, see <http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html.ncbi/barcode/>

throughout the world. DNA barcodes therefore provide the basis of a promising alternative strategy because they require only collection of DNA and not the immediate taxonomic identification of specimens. Although barcodes have many other uses, e.g., identification of novel species, taxonomic classification, and phylogeny, their application to cataloguing biodiversity justifies restricting this chapter to the measurement of a barcode's efficacy in identifying known species.

In its essence, a barcode is any standardized subset of DNA from a taxonomic specimen (1, 2). The subset may vary, depending on readily recognizable features of a specimen (e.g., is the specimen a vertebrate? a plant? an insect? etc.). If computers could identify the species of a specimen from its barcode, then the barcode would provide a database key for retrieving taxonomic information pertinent to the specimen. A computer catalog of species on Earth then becomes a technical possibility. Early studies indicated that the sequence of cytochrome c oxidase 1 (CO1) gene could correctly identify many species (3), so selection of CO1 as a primary barcode followed naturally (4–10).

Although the selection of a DNA barcode has been natural for some species, it has been problematic for others, particularly plants (11–14) and insects (15, 16). The lack of a clear consensus for a barcode in those species has stimulated interest in the objective, quantitative measurement of the efficacy of a barcode in identifying species. Consensus on an actual barcode for some species remains tentative, but nonetheless, a consensus on measuring barcode efficacy has emerged (14, 15, 17). This chapter summarizes the consensus and indicates how to construct studies to evaluate the relative merits of competing barcodes. For practical methods, the reader is invited to view <http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html.ncbi/barcode/>, a Web site providing information on computer programs pertinent to barcodes. Web pages are supposed to be self-explanatory, so to avoid undue brevity, the second section in this chapter provides some rationale for the computer programs for evaluating barcodes. The third section provides a practical summary of the entire chapter.

2. The Measurement of the Efficacy of Species Identification

To fix our terminology, the term “marker” connotes any contiguous region of DNA (coding or non-coding), whereas the term “barcode” connotes the aggregate of the one or more markers in the “standardized subset of DNA” referred to in the Introduction. Presently, all barcode markers are marker genes like CO1, matK, etc.

In slowly evolving organisms like plants, however, intergenic spacers (DNA regions flanked by two genes) are still worthy of consideration as potential markers, because they usually diverge faster than genes, while their ends are still conserved, providing primers for PCR (17, 18). As described below, however, multiple sequence alignments (MSAs) of intergenic markers might complicate the workflow in a barcode database.

To have practical meaning, any measurement of the efficacy of species identification must mirror the performance of a database based on the prospective barcode. In practice, users query the database with a barcode retrieved from a specimen; the database returns the species identification as output, with the assignment “unknown” for any species apparently not yet in the database. Because this chapter restricts itself to discussing the identification of known species, it assumes that each query to the barcode database represents a specimen belonging to a species already in the database.

2.1. The Database

The first step in estimating the efficacy of several prospective barcodes is to assemble the corresponding databases. To ensure the proper controls, specimens not having sequences in every marker database should be eliminated from consideration (14), because if the databases do not contain exactly the same specimens, there might be unappreciated but influential biases. Consider, e.g., a hypothetical experiment that extracts from GenBank all sequences corresponding to two prospective markers, Marker A and Marker B. If Marker A has been the default marker of choice, whereas Marker B has been considered as the last hope for resolving species after Marker A has failed, the GenBank entries for Marker B might be biased toward a subset of particularly difficult specimens. Thus, on GenBank data, Marker B might have fewer correct species assignments than Marker A, even though Marker B is in fact better at resolving species than Marker A. Moreover, relative to a barcode database, GenBank taxonomy is undependable, and undependable taxonomy improperly influences conclusions by occasionally penalizing correct species identification. In addition, GenBank entries do not usually identify individual taxonomic specimens. GenBank data are therefore particularly unsuited to studying barcodes based on more than one marker, because the sequences from different markers cannot be associated with a single specimen. Although studies based on GenBank data have obvious scientific interest, they do not have the same status as a controlled taxonomic study. In summary, the choice of database affects conclusions, so care must be taken that the database reflects the scientific aims of a study.

Figure 1 shows some pertinent results for *trnH-psbA*, a potential barcode marker in plants. By using pairwise alignment and various evolutionary distances in the procedures described below, the best overall probability of correct identification (PCI) in Fig. 1 is about 0.50, which is noticeably lower than the overall PCI of

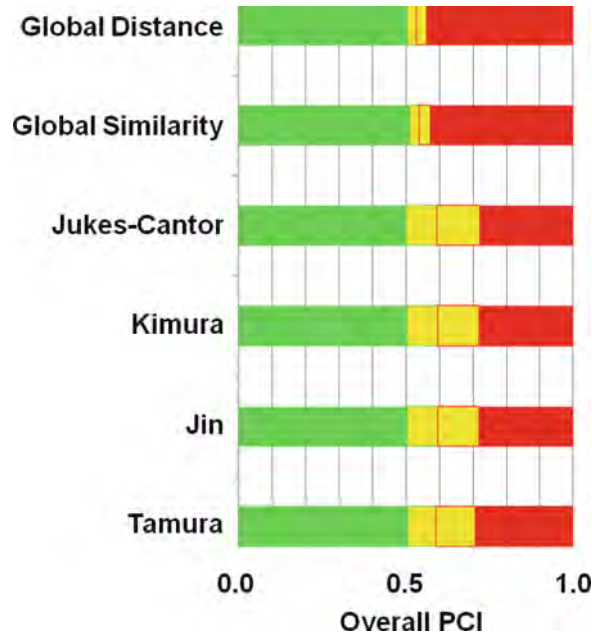


Fig. 1. Overall PCIs for *trnH-psbA*. Figure 1 graphs the overall PCI (on the X-axis) from assigning plant species with *trnH-psbA* sequences collected from GenBank. (The corresponding FASTA file can be obtained at http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html_ncbi/html/bib/116.html). Assignment used a nearest neighbor algorithm and one of six separations (on the Y-axis). The six separations were: (1) Global Distance; (2) Global Similarity; and four evolutionary distances: (3) Jukes-Cantor (38); (4) Kimura (2-Parameter) (39); (5) Jin (using a gamma distribution with parameter 1) (40); and (6) Tamura (41). The pairwise sequence alignment used either the HOX70 scoring matrix

$$\begin{vmatrix} & A & C & G & G \\ A & 91 & -114 & -31 & -123 \\ C & -114 & 100 & -125 & -31 \\ G & -31 & -125 & 100 & -114 \\ T & -123 & -31 & -114 & 91 \end{vmatrix},$$

with a gap of length k receiving a penalty $\Delta(k)=400+30k$, or the NCBI DNA scoring system (1 for a match, -3 for a mismatch, with a gap of length k receiving a penalty $\Delta(k)=5+2k$). Perhaps surprisingly, the overall PCIs for the two scoring systems were visually indistinguishable. Global Distance is the global alignment score; Global Similarity is the actual global alignment score divided by the maximum possible global alignment score for sequences of the same length (42). The *green part of the horizontal bars* gives the unambiguously correct fraction of species assignments, where every specimen had as nearest neighbors only specimens from the same species; the *yellow part*, the ambiguously correct fraction where every specimen had as nearest neighbors specimens a mix from both the same and other species (with the *red border* indicating the average fraction of the ambiguously correct fraction matching specimens from different species); and the *red part*, the unambiguously incorrect fraction where every specimen had only nearest neighbor specimens from other species.

0.69 from a controlled taxonomic study (14), suggesting that the GenBank entries for *trnH-psbA* might contain biases, relative to a controlled taxonomic study. The corresponding FASTA sequence file (see the Supplementary Materials) in fact contained genetic crosses (denoted by “x”) and tentative species assignments (denoted by “sp.”, “cf.”, “aff.”), which were obscure, until the Web tools mentioned above found them.

2.2. Species Assignment Algorithm

Once an appropriate database has been selected, the computer must assign a species to each barcode query (or declare its failure to assign). The next step, therefore, is to select a computer algorithm for assigning each specimen and its barcode sequence to a species. No algorithm seems to improve noticeably on assigning to a specimen the species of its nearest neighbor within a barcode database (19, 20). Thus, many algorithms begin by estimating a “separation” between the barcode sequences in two specimens. (The term “separation” is preferable to “distance”, which connotes some specific mathematical properties not necessary to barcodes.)

Separation can be based on: (1) sequence alignment similarities, (2) sequence alignment distances, (3) evolutionary distances (which usually require prior alignment of the barcode sequences), or (4) alignment-free distances. Studies have compared different measures of separation, but they are too limited to draw definitive conclusions about which separation provides the best species assignments. There are, however, some distinctly bad measures of separation.

Like any assignment method, species assignment should use all available information. BLAST is a popular sequence comparison tool (21, 22), but as a measure of separation it can mislead, because it compares two sequences with local alignment, which matches and scores only the two most similar subsequences within two sequences (see Fig. 2, which diagrams some of the differences between local and global alignments). Global alignment, which matches the entire length of sequences, is better for measuring the separation of barcode marker sequences. In intergenic markers particularly, BLAST has the possible weakness of matching only small subsequences, because alignments within intergenic spacers often contain large gaps. Short subsequences can exhibit convergent evolution (homoplasy) (23), so on the one hand a BLAST local alignment might make distant species appear spuriously close. On the other hand, a global alignment might resolve the species by highlighting dissimilarities across the whole marker. In the context of barcodes, therefore, a global alignment (e.g., with some close relative of the Needleman–Wunsch Algorithm (24)) is generally preferable to a local alignment (e.g., with the Smith–Waterman Algorithm (25) or BLAST). Other types of alignments exist, but there is little reason to expect them to assign species notably better than global alignment.

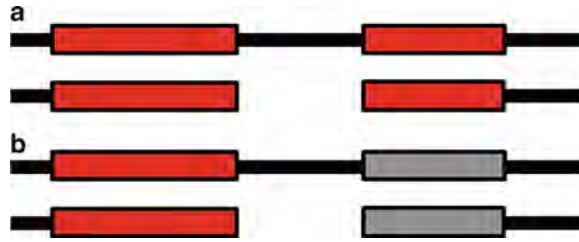


Fig. 2. Two types of alignment, global and local. (a) shows a global alignment of two sequences (*black lines*). Global alignment is an alignment along the complete length of the sequences, so it bridges a gap in the second sequence (*white space*), to include all pairs of similar subsequences (*red rectangles*). (b) shows a local alignment of the same two sequences. Local alignment aligns only the pair of most similar subsequences in the sequences, so it does not bridge the gap in the second sequence and does not include the smaller subsequence alignment (*now shown in gray*). Local alignment can be misleading when identifying species with barcodes because it does not incorporate all available sequence information.

MSAs might be more problematic for intergenic markers than for marker genes like COI, because intergenic MSAs usually contain many gaps, disrupting the alignment columns representing evolutionary relationships. In practice, the Barcode of Life Database (<http://www.boldsystems.org>) stores sequences in a global MSA, by using the program HMMer (26) to align sequences before comparing the corresponding barcode marker genes. In fact, many publicly available tools (e.g., MUSCLE (27) or MAFFT (28)) could create barcode MSAs interchangeably with HMMer. The point of using MSAs in a large barcode database, however, is that MSA can be much faster than pairwise sequence alignment. (If there are N barcodes in a database, pairwise alignment requires time proportional to N^2 .) Although bioinformatics should adapt to the needs of biology and not vice versa, the selection of an intergenic marker as a barcode might exclude MSAs in the workflow of large barcode databases, causing awkward (but probably not insuperable) difficulties.

As separations, the relative merits of global alignment similarity, global alignment distances, or evolutionary distances based on a global alignment have not yet been clearly established, although the differences in species assignment are probably small. Alignment distances and similarities model insertions and deletions in sequences, which are not as well understood as nucleotide substitutions used in evolutionary distances. As a separation, p-distance (the proportion p of alignment pairs containing differing nucleotides) is particularly simple and well-known to taxonomists (20), but in fact no separation based on global alignment has shown any clear superiority in species assignment over the others.

Other species assignment algorithms should be mentioned (29, 30). Many probabilistic algorithms, in particular those producing phylogenetic trees (31, 32), are now a commonplace in taxonomy.

Unfortunately, most probabilistic computations are much slower than the nearest neighbor algorithms above. Because they do not noticeably improve identification, they have not found a place in automatic species identification. Alignment-free algorithms are simple and provide faster computation than alignment-based methods (20, 33), but presently, they have not been widely adopted in species identification.

2.3. Probability of Correct Identification

With an appropriate database and species assignment algorithm in hand, a scientist interested in barcode efficacy must measure the algorithm's success in identifying species. Any reasonable measure of barcode efficacy should reflect the probability that a database based on the prospective barcode identifies a specimen's species correctly. Consensus has therefore emerged on "the probability of correct identification" (PCI) as the appropriate measurement of barcode efficacy (14, 15, 17). The ambiguities in the definition of PCI accommodate legitimate scientific disagreement about success in species identification, so the concept of PCI actually embraces a broad class of measures.

Consider a particular data set, and assume that PCI can be defined for each species within the data set. The overall PCI for the data set is the average of the species PCIs, taken over all species in the data set. If a few data subsets are particularly important (e.g., angiosperm, basal, and gymnosperm subsets within a plant data set), the PCI for the subsets can be reported separately. In principle, the PCI for each species could be weighted to reflect the species' importance or the number of specimens representing it in the data set. In practice, however, scientists have not weighted averages when calculating overall PCI. Thus, to calculate the overall PCI of a data set, we now require only a species PCI, a probability to quantify success in identifying each fixed species.

To calculate a species PCI, one can perform a leave-one-out procedure, sometimes called "the jackknife" in statistics (34). Remove each specimen in a species in turn from the database, and consider the separation of the removed specimen from the specimens of the same species remaining in the database. (The leave-one-out procedure cannot sensibly be applied if a species has only a single specimen in the database. Because a singleton species must therefore be omitted from the average in the overall PCI, it usually represents wasted experimental effort. It does, however, provide a "decoy," which provides a realistic impediment to correct species assignment.)

Scientists legitimately disagree over the definition of "success" in species identification. Some scientists might consider "success" theoretically, as a monophyly, where every specimen in the species is closer to all specimens in the species than to any other specimen (14). On success, the species PCI is 1; on failure, it is 0. Other scientists might consider success more pragmatically, as a correct assignment of the species, where each specimen in the species

has as its nearest neighbor(s) only specimens in the species (15). Again, if so, the species PCI is 1; if not, it is 0. The following additional conditions can contribute to success or failure, as desired: ties outside the species for a nearest neighbor, assignment of specimens from other species to the species in question, etc.

Some authors have advanced less stringent criteria for success (e.g., for $k > 1$, the specimen's nearest neighbors must contain at least one other specimen from the same species) (33). The species PCI has also been calculated as the fraction of specimens within a species whose nearest neighbor gives the correct assignment (17). Any specific choice might be appropriate in different circumstances, depending on the scientific aim.

Some authors experimented with placing additional conditions on “success” as defined above, e.g., sequence difference (p-distance) thresholds, such as 2% or 3% (15). Detection of unknown species with sequence identity thresholds seems artificial, however (35). The notion of “species” could be redefined by DNA thresholds (1, 2, 36, 37), but such redefinitions generate many conflicts with traditional taxonomy (15).

2.4. PCR Failure

PCI should estimate the success in correctly identifying a known species. Under present technology, species identification with a DNA barcode requires the following criteria:

1. At least part of the barcode sequence must be present in the specimen.
2. Laboratory procedures must physically extract it from the specimen.
3. PCR primers must amplify it.
4. It must be sequenced.
5. It must diverge sufficiently, to distinguish species.
6. It must not diverge excessively, so specimens from a single species remain similar and identifiable.

Thus, PCI must account for PCR failure, if it is to estimate identification success under present technology. Recall that the overall PCI is the average of the PCI for each individual species. The Appendix discusses PCR failure for a barcode based on several markers. For simplicity, this subsection considers here only a barcode based on a single marker. We revise the species PCI to account for PCR failure, as follows. According to the procedures in the preceding subsection (which ignore PCR failure), let the species have PCI p ; and let s be the fraction of specimens from the species with a successful PCR. (Note that s is estimated from all specimens, whereas p is estimated solely from specimens with a successful PCR.) A reasonable procedure might average the “PCR-adjusted species PCI” $p' = ps$ over all species to produce a

“PCR-adjusted overall PCI.” The PCR-adjusted overall PCI faithfully reflects the efficacy of species identification with present technology, whereas the overall PCI (which ignores specimens where PCR failed) reflects the efficacy of species identification with a perfect PCR technology.

Technology reduces PCR failure rates, so arguments have been advanced that PCR failure should be ignored (14). The PCI after any technological advance, however, is bounded below by the PCR-adjusted overall PCI (which reflects present PCR technology); similarly, it is bounded above by the overall PCI (which ignores specimens with failed PCR). The bounds demonstrate that technological advance by itself does not preclude a sober assessment of future prospects. Like any numerical result from a definite procedure with a sensible meaning, the PCR-adjusted overall PCI is useful, and its deliberate omission merely undermines rational discussion about the relative merits of potential barcodes.

2.5. Statistical Sampling Error

The overall PCI is the (unweighted) average of the species PCIs. Let us make a reasonable approximation that species PCIs are mutually independent across all species. Any database is a sample of all possible species, so the overall PCI from the database is an estimate of the “true” overall PCI p . As such, it has a sampling error, calculable with the binomial distribution. Let n be the number of species contributing to the overall PCI. Under mild assumptions (given below), a binomial estimate \hat{p} is normally distributed with mean p and standard deviation $\sqrt{p(1-p)/n}$. Thus, the confidence

interval $\left[\hat{p} - z\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z\sqrt{\hat{p}(1-\hat{p})/n} \right]$ contains the true overall PCI p with a confidence determined by z in conjunction with the normal distribution. The larger z is, the broader the interval becomes, and the greater the probability that the interval contains the true value of p . As approximate examples, $z = 2$ yields an 95% confidence interval; $z = 2.6$, 99%, etc. (As a useful rule of thumb, the normal approximation holds, if $n \geq 20$ and the confidence interval does not include 0.0 or 1.0.) Confidence intervals are worth calculating, because they are often surprisingly broad.

As an aside, the confidence intervals for the overall PCI are crucial to evaluating the relative merits of tentative barcodes, but they have little direct bearing on one’s confidence in the species assignment of a specific specimen, for the following reason. Most taxonomists probably prefer a barcode for which assignment errors are confined to a few species, rather than to have the same errors spread across many species. (If nothing else, alternative strategies might be available for assigning a small number of problematic species.) Overall PCI faithfully reflects taxonomists’ barcode preferences, but the evaluation of a specific species assignment poses a different problem, requiring a different solution.

3. The Summary of the Workflow

Selection of a DNA barcode has been problematic for some species, but there is now a general consensus on the measurement of barcode efficacy. The procedure for measuring barcode efficacy can be broken into several steps.

First, assemble databases corresponding to the prospective barcodes. The choice of database must be given careful consideration because it can noticeably influence a study's conclusions. To ensure proper controls, specimens not having a sequence in every marker database should be eliminated from consideration. Because GenBank taxonomy might be undependable, and because most GenBank sequences do not specify a corresponding taxonomic specimen, studies based on GenBank data do not have the same status as a controlled taxonomic study, particularly for barcodes based on more than one marker.

Second, select a computer algorithm for assigning species to barcode sequences. No algorithm seems to improve noticeably on assigning to a specimen the species of its nearest neighbor within a barcode database. A global alignment (e.g., with Needleman–Wunsch algorithm, or some similar algorithm) is recommended, to take advantage of all the information in a barcode sequence. By contrast, BLAST is a local alignment program, which might match only small subsequences within two sequences. Thus, the use of BLAST runs an unnecessary risk when evaluating any prospective barcode, particularly one with an intergenic marker. As long as alignments are in essence global, alignment similarities, alignment distances, and evolutionary distances like *p*-distance, Kimura 2-Parameter Distance, etc., seem to have approximately equal efficacies in identifying species.

Consensus has emerged on “the probability of correct identification” (PCI) as the appropriate measurement of barcode efficacy. The overall PCI for a data set is the average of the species PCIs, taken over all species in the data set. If a few data subsets are particularly important (e.g., angiosperm, basal, and gymnosperm subsets within a plant data set), the PCI for the subsets can be reported separately.

To calculate a species PCI, remove in turn each specimen in the species from the database, and consider its separation from the remaining specimens (under, e.g., *p*-distance). Various definitions of identification success within a species are possible: (1) every specimen in the species is closer to all other specimens in the species than to any other specimen; (2) each specimen in the species has another specimen in the species as its nearest neighbor; (3) more stringent versions of the two foregoing definitions, where ties outside the species for a nearest neighbor, or assignment of other species to the species in question, also connote failure; (4) less stringent criteria for success (e.g., for $k > 1$, the specimen's nearest

k neighbors must contain at least one other specimen from the same species; or (5) probabilistic measures of success, like the fraction of specimens within a species displaying one of the foregoing definitions of success. Scientific purpose makes different definitions of “successful assignment” appropriate to different circumstances.

To estimate success under present technology, PCI must account for PCR failure. Although the case of a barcode with several markers has been relegated to the [Appendix](#), the case of a barcode with only one marker poses no difficulties. Simply estimate the rate of PCR failure within each species by using all specimens, not just the ones with completely successful PCRs. Multiplication of a species PCI by the PCR success rate within the species yields a “PCR-adjusted” species PCI, which can then be averaged over species to yield a PCR-adjusted overall PCI. The overall PCI after technological advance is bounded below by the PCR-adjusted overall PCI; similarly, it is bounded above the overall PCI (which derives from PCR successes only). Thus, present technology bounds prospects for an overall PCI.

A database provides a statistical sample of all possible data. The overall PCI calculated from a database is therefore a statistical estimate of the true overall PCI, and as such, it yields an estimate with a statistical error. The errors are sometimes surprisingly large, and the differences in barcode efficaciousness correspondingly small.

For software relevant to this chapter, see <http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html.ncbi/barcode/>.

Acknowledgment

This research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI.

Appendix

For a barcode with several markers, each of which can have a failed PCR, specimen identification ultimately relies on the markers with a successful PCR. To quantify the identification process, number the markers $\{1, 2, \dots, m\}$, and consider any subset M of $\{1, 2, \dots, m\}$. For a particular specimen, let the probability that M is the subset of markers with PCR success be denoted by s_M , and let the PCI for the barcode based on the marker subset M be p_M . A species PCI p can then be calculated from the values of s_M and p_M (although the calculation depends on the definition of species PCI: see Section 2.3 for various definitions.)

One very reasonable definition of the PCR-adjusted species PCI is the average $p = \sum_{(M)} p_M s_M$. For the case of a barcode based on a single marker, e.g., M is a subset of $\{1\}$, i.e., the empty set $\{\}$ or $\{1\}$. Because the empty set $\{\}$ corresponds to a complete absence of information about a specimen, the corresponding PCI is $p_{\{\}} = 0$, so $p = p_{\{\}} s_{\{\}} + p_{\{1\}} s_{\{1\}} = p_{\{1\}} s_{\{1\}}$, which agrees with the formula for the PCR-adjusted PCI in the main text, for a barcode based on a single marker.

References

1. Hebert PD, Cywinska A, Ball SL, Dewaard JR (2003) Biological identifications through DNA barcodes. *Proc Biol Sci* 270:313–321
2. Floyd R, Abebe E, Papert A, Blaxter M (2002) Molecular barcodes for soil nematode identification. *Mol Ecol* 11:839–850
3. Hebert PD, Ratnasingham S, Dewaard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Biol Sci* 270:S96–S99
4. Hajibabaei M, Janzen DM, Burns JM et al (2006) DNA barcodes distinguish species of tropical lepidoptera. *Proc Natl Acad Sci U S A* 103:968–971
5. Hogg ID, Hebert PDN (2004) Biological identification of springtails (hexapoda: Collembola) from the canadian arctic, using mitochondrial DNA barcodes. *Can J Zool* 82: 749–754
6. Lorenz JG, Jackson WE, Beck JC, Hanner R (2005) The problems and promise of DNA barcodes for species diagnosis of primate biomaterials. *Philos Trans R Soc Lond B Biol Sci* 360:1869–1877
7. Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol* 3:e422
8. Saunders GW (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philos Trans R Soc Lond B Biol Sci* 360:1879–1888
9. Smith MA, Fisher BL, Hebert PDN (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philos Trans R Soc Lond B Biol Sci* 360:1825–1834
10. Smith MA, Woodley NE, Janzen DH et al (2006) DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (diptera: Tachinidae). *Proc Natl Acad Sci U S A* 103: 3657–3662
11. Chase MW, Salamin N, Wilkinson M et al (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philos Trans R Soc Lond B Biol Sci* 360:1889–1895
12. Cowan RS, Chase MW, Kress JW, Savolainen V (2006) 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. *Taxon* 55:611–616
13. Kress WJ, Erickson DL (2008) DNA barcodes: genes, genomics, and bioinformatics. *Proc Natl Acad Sci U S A* 105:2761–2762
14. Cbol Plant Working Group (2009) A DNA barcode for land plants. *Proc Natl Acad Sci U S A* 106:12794–12797
15. Meier R, Shiyang K, Vaidya G, Ng PK (2006) DNA barcoding and taxonomy in diptera: a tale of high intraspecific variability and low identification success. *Syst Biol* 55:715–728
16. Huang D, Meier R, Todd PA, Chou LM (2008) Slow mitochondrial *coI* sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. *J Mol Evol* 66:167–174
17. Erickson DL, Spouge JL, Resch A et al (2008) DNA barcoding in land plants: developing standards to quantify and maximize success. *Taxon* 13:1304–1316
18. Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding *rbcl* gene complements the non-coding *trnH-psbA* spacer region. *PLoS One* 2:e508
19. Austerlitz F (2007) Comparing phylogenetic and statistical classification methods for DNA barcoding. Paper presented at the second international barcode of life conference, Taipei, Taiwan, 2007
20. Little DP, Stevenson DW (2007) A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. *Cladistics* 23:1–27
21. Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
22. Altschul S (1999) Hot papers – bioinformatics – gapped blast and psi-blast: a new generation

- of protein database search programs by s.F. Altschul, t.L. Madden, a.A. Schaffer, j.H. Zhang, z. Zhang, w. Miller, d.J. Lipman – comments. *Scientist* 13:15
23. Wouters MA, Husain A (2001) Changes in zinc ligation promote remodeling of the active site in the zinc hydrolase superfamily. *J Mol Biol* 314:1191–1207
 24. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:443–453
 25. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147:195–197
 26. Eddy SR (1995) Multiple alignment using hidden markov models. *Proc Int Conf Intell Syst Mol Biol* 3:114–120
 27. Edgar RC (2004) Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113
 28. Katoh K, Misawa K, Kuma K, Miyata T (2002) Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res* 30:3059–3066
 29. Matz MV, Nielsen R (2005) A likelihood ratio test for species membership based on DNA sequence data. *Philos Trans R Soc Lond B Biol Sci* 360:1969–1974
 30. Nielsen R, Matz M (2006) Statistical approaches for DNA barcoding. *Syst Biol* 55: 162–169
 31. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17:368–376
 32. Felsenstein J (1988) Phylogenies from molecular sequences – inference and reliability. *Annu Rev Genet* 22:521–565
 33. Kuksa P, Pavlovic V (2009) Efficient alignment-free DNA barcode analytics. *BMC Bioinformatics* 10:S9
 34. Efron B, Stein C (1981) The jackknife estimate of variance. *Ann Stat* 9:586–596
 35. Ferguson JWH (2002) On the use of genetic divergence for identifying species. *Biol J Linnean Soc* 75:509–516
 36. Blaxter M, Mann J, Chapman T et al (2005) Defining operational taxonomic units using DNA barcode data. *Philos Trans R Soc Lond B Biol Sci* 360:1935–1943
 37. Lambert DM, Baker A, Huynen L et al (2005) Is a large-scale DNA-based inventory of ancient life possible? *J Hered* 96(3):279–284
 38. Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) *Mammalian protein metabolism*. Academic, New York, pp 21–123
 39. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16:111–120
 40. Jin L, Nei M (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol Biol Evol* 7:82–102
 41. Tamura K (1994) Model selection in the estimation of the number of nucleotide substitutions. *Mol Biol Evol* 11:154–157
 42. Waterman MS, Smith TF, Beyer WA (1976) Some biological sequence metrics. *Adv Math* 20:367–387

RESEARCH ARTICLE

Open Access

The *Physalis peruviana* leaf transcriptome: assembly, annotation and gene model prediction

Gina A Garzón-Martínez^{1†}, Z Iris Zhu^{2†}, David Landsman², Luz S Barrero^{1,3} and Leonardo Mariño-Ramírez^{1,2,3*}

Abstract

Background: *Physalis peruviana* commonly known as Cape gooseberry is a member of the Solanaceae family that has an increasing popularity due to its nutritional and medicinal values. A broad range of genomic tools is available for other Solanaceae, including tomato and potato. However, limited genomic resources are currently available for Cape gooseberry.

Results: We report the generation of a total of 652,614 *P. peruviana* Expressed Sequence Tags (ESTs), using 454 GS FLX Titanium technology. ESTs, with an average length of 371 bp, were obtained from a normalized leaf cDNA library prepared using a Colombian commercial variety. *De novo* assembling was performed to generate a collection of 24,014 isotigs and 110,921 singletons, with an average length of 1,638 bp and 354 bp, respectively. Functional annotation was performed using NCBI's BLAST tools and Blast2GO, which identified putative functions for 21,191 assembled sequences, including gene families involved in all the major biological processes and molecular functions as well as defense response and amino acid metabolism pathways. Gene model predictions in *P. peruviana* were obtained by using the genomes of *Solanum lycopersicum* (tomato) and *Solanum tuberosum* (potato). We predict 9,436 *P. peruviana* sequences with multiple-exon models and conserved intron positions with respect to the potato and tomato genomes. Additionally, to study species diversity we developed 5,971 SSR markers from assembled ESTs.

Conclusions: We present the first comprehensive analysis of the *Physalis peruviana* leaf transcriptome, which will provide valuable resources for development of genetic tools in the species. Assembled transcripts with gene models could serve as potential candidates for marker discovery with a variety of applications including: functional diversity, conservation and improvement to increase productivity and fruit quality. *P. peruviana* was estimated to be phylogenetically branched out before the divergence of five other Solanaceae family members, *S. lycopersicum*, *S. tuberosum*, *Capsicum spp*, *S. melongena* and *Petunia spp*.

Keywords: *P. peruviana*, Solanaceae, ESTs, Functional annotation, Gene model, Phylogenetics

Background

Physalis peruviana, also known as Cape gooseberry is a tropical fruit from the Solanaceae family, which includes many agriculturally important crops including potato, tomato, pepper, eggplant and tobacco [1]. The Cape gooseberry fruit contains high levels of vitamin A, C and B-complex, as well as compounds of anti-inflammatory and antioxidant properties [2]. Supercritical carbon

dioxide extracts of *P. peruviana* leaves were shown to induce cell cycle arrest and apoptosis in human lung cancer H661 cells [3]. Recently, 4β-Hydroxywithanolide (4βHWE) isolated from *P. peruviana* aerial parts (stems and leaves) was demonstrated to be a potential DNA-damaging and chemotherapeutic agent against lung cancer [4]. In Colombia, this fruit has become promissory with high demand in European markets, mainly due to its unique taste, attractive color and shape as well as its potential health value. *P. peruviana* is a source of health related compounds found in the fruit and other parts of the plant including leaves and stems. Despite its nutritional and medical importance, current absence of *P. peruviana* genetic and genomic resources makes in-depth molecular studies on the plant difficult. Until this study, there were only a few partial *P. peruviana* gene

* Correspondence: marino@ncbi.nlm.nih.gov

†Equal contributors

¹Plant Molecular Genetics Laboratory, Center of Biotechnology and Bioindustry (CBB), Colombian Corporation for Agricultural Research (CORPOICA), Bogota, Colombia

²Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, United States of America, Bethesda, MD, USA

Full list of author information is available at the end of the article

sequences in public databases, mainly as a result of phylogenetic studies in the Solanaceae family [5,6]. Therefore, there is a pressing need for efforts to obtain global genetic and genomic information from the Cape gooseberry, *P. peruviana*.

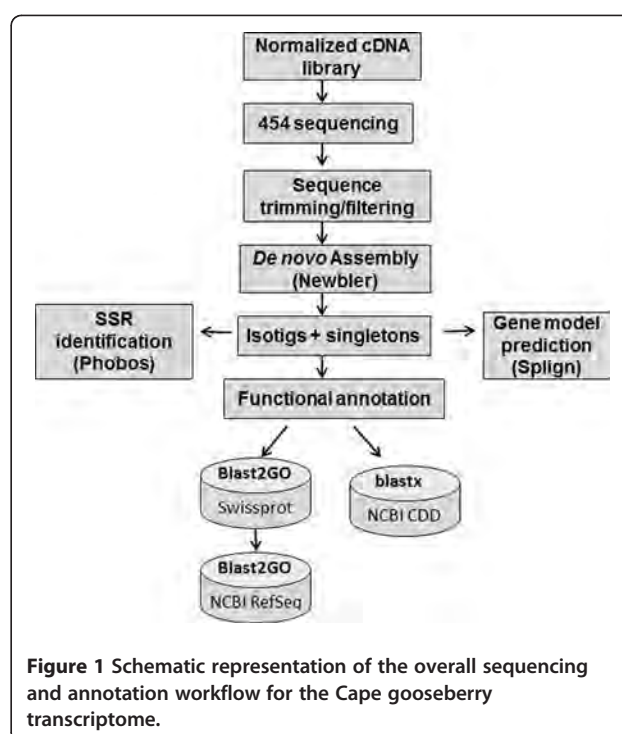
Advances in next generation sequencing (NGS) technology over the past few years have made it possible to rapidly perform *de novo* transcriptome and even genome assembly for non-model organisms with no or little prior genomic data available [7]. However, polyploidy and the large size of many plant genomes, which is predominantly due to amplification of repetitive elements or sometimes partial genome duplication [8], pose challenges to *de novo* whole genome assembly of plants. As such, EST sequencing, which avoids non-coding and repetitive DNA components, is a cost-effective and commonly used strategy to analyze the transcribed portion of a genome. Availability of ESTs represent a valuable resource for research as they provide comprehensive information regarding the transcriptome facilitating gene discovery and genome annotation and aiding in the determination of phylogenetic relationships [9]. An increasing number of successful studies have been published describing EST sequencing and *de novo* transcriptome assembly for large-scale gene discovery [9-18].

Here we describe the sequencing and assembly of the first *P. peruviana* leaf transcriptome from its cDNA-derived ESTs using the 454 GS-FLX Titanium technology, as well as *in silico* functional annotation and gene model prediction of the assembled transcriptome. The overall workflow of the project is represented in Figure 1. This first transcriptome draft will provide valuable resources for the development of molecular genetic tools that can be used in agronomic trait related marker discoveries, in addition to studies that aim to solve phytosanitary, fruit quality and production problems.

Results and discussion

EST sequencing and assembly

We performed three fourths 454 GS FLX Titanium run on one normalized cDNA library constructed from *P. peruviana* leaf tissue, generating approximately 336 Mbp of sequence data from 652,614 reads with an average length of 375 bp (Figure 2). After a trimming process by SeqClean [19], which removes adaptors, primer sequences, poly-A tails, as well as short, longer and low quality sequences, a total of 641,512 high quality reads were obtained with an average length of 371 bp. *De novo* transcriptome assembly was performed using Newbler 2.5.3 [20], which has been shown to perform better than a number of other commonly used assemblers [21]. Table 1 shows the transcriptome sequencing and assembly statistics, 79.66% of the reads were assembled into 29,911 contigs, and then further into 24,014 isotigs, with



an average assembled length of 1,638 bp. The isotig N50 length was 2,504 bp. All isotigs that share common contigs belong to the same isogroup, presumably equivalent to one gene locus containing multiple alternatively spliced transcripts. The 24,014 assembled isotigs are part of 14,049 isogroups (equivalent to an average 1.7 transcripts per gene), among which 9,655 isogroups have only one isotig each. Isotigs whose length exceeded 200 bp (23,964 in total) were kept for further analysis. The remaining 20.34% reads are singletons, which cannot be connected with any other reads. The 110,921 singletons were kept for further analysis. The average coverage of assembled isotigs is estimated to be 9.1X. The number goes down to 3.9X if we include all the singletons as the effective transcribed portion. Isotigs and singletons together will be referred as cDNAs in the rest of the manuscript. The raw data files are available at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) accession number SRP005904. The assembled reads were deposited in the Transcriptome Shotgun Assembly (TSA) Database [GenBank:JO124085-JO157957].

Functional annotation

As the first step for assigning putative functions to the *P. peruviana* transcriptome, BLASTX searches [22,23] were used to align the cDNAs to the UniProtKB/Swiss-Prot and NCBI RefSeq databases. A total of 19,162 isotigs and 35,428 singletons had a BLAST hit (with an

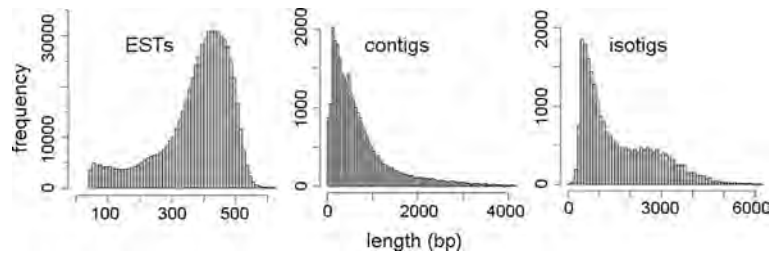


Figure 2 Length distributions of *P. peruviana* EST reads (left), assembled contigs (center) and isotigs (right). Data obtained after sequencing with three fourths run of 454 GS FLX Titanium of the normalized leaf cDNA library.

expectation value $< 1e-5$) to known proteins and matched 8,721 and 15,192 unique protein accessions, respectively. More than 99% of the BLASTX hits from both isotigs and singletons were from plant proteins. Compared to isotigs, a much greater percentage of singletons do not have any significant hits (68%), which could be mainly due to their short lengths. Using Blast2GO [24], we retrieved gene ontology (GO) terms and enzyme commission numbers (EC) for the *P. peruviana* cDNAs (Table 2) from the BLASTX output described above. A total of 33,105 GO terms were assigned to 12,672 cDNAs (including isotigs and singletons). Among all the GO terms extracted, 13,935 (42%) belong to the Molecular Function class, 10,375 (31%) to Biological Process class and 8,795 (27%) to Cellular Component class. There are 7,519 cDNAs assigned to multiple GO terms.

The biological process (BP) GO category comprise different types of metabolic processes which are the most represented categories: there are 4,620 sequences associated with metabolic processes (GO level 2), which is expected, since the metabolic network in plants is by far more extensive compared to other organisms [25]. We found GO terms associated with primary metabolites, which include the universal building blocks of sugars, amino acids, nucleotides, lipids, and energy sources that are essential for plant survival. Additionally, we found GO terms associated with secondary metabolites that play key roles in maintaining plant fitness including ones that function in the protection of plants against microbial, viral infections and UV radiation. Shown in Figure 3 are a number of GO terms (BP category, level 4) that are

abundant and relevant to plant physiology, like the metabolic processes of nitrogen compound, nucleotide, carbohydrate, amine and phosphorus. Another category worthy to mention is “response to stimulus” (BP category, level 2). We found 1,120 sequences associated with this category, which include candidate genes for resistance to pathogen attacks. Shown in Figure 3 are a number of level 4 GO categories including: response to organic substance, defense response and response to hormone stimulus. In the molecular function (MF) category, 30% of the *P. peruviana* cDNAs have high similarity to proteins with transferase or hydrolase activity (GO level 3) that includes genes associated with secondary metabolic synthesis pathways [9,10]. Other abundant level 3 MF categories include: nucleotide binding, ion binding and oxidoreductase binding (Figure 3).

We were able to assign 129 unique enzyme commission (EC) numbers to 1,671 *P. peruviana* cDNAs, where 25 unique EC numbers were in turn assigned to 52 metabolic pathways linked to 1,255 cDNAs (Table 3). We found 187 cDNAs involved in thiamine metabolism in addition to 84 sequences associated with secondary metabolite biosynthesis and 53 assigned to the phenylpropanoid biosynthesis pathway. These pathways are of particular interest in *Physalis* as thiamine has been known to induce defense response in plants through the salicylic acid and Ca^{2+} -related signaling pathways [26,27] and may play roles in biotic or abiotic stress [28]. Furthermore, secondary metabolites such as phenylpropanoids play important roles in resistance mechanisms to pathogens and recently have also been used in medicinal applications including antioxidants, anticancer and anti-inflammatories [2,28].

Table 1 *P. peruviana* transcriptome assembly overview

	Filtered EST reads	Contigs	Isotigs	Isogroups	Singletons
	641,512	29,911	24,014	14,049	110,921
Average length (bp)	371	743	1,638		354
N50 size (bp)		1,438	2,504		

Protein domains encoded by the *P. Peruviana* leaf transcriptome

A total of 12,974 *P. peruviana* cDNAs were found to have significant similarities to 3,117 protein domains present in the NCBI CDD (Conserved Domain Database) [30]. The most abundant domain present in proteins encoded by the *P. peruviana* transcriptome is the

Table 2 *P. peruviana* transcriptome functional annotation overview

	Isotigs	Singletons	Total
Sequences with BLAST hits	19,162	35,428	54,590
Sequences annotated with GO terms	4,915	7,757	12,672
GO Terms associated with the sequences	12,675	20,430	33,105
Sequences associated with EC numbers	601	1,070	1,671

pentatricopeptide repeat domain (PPR), found in 350 cDNAs. The PPR containing proteins are commonly found in the plant kingdom and although its function is still unclear, the PPR domain has been found in proteins involved in RNA editing in a number of recent studies [31-34]. Following the PPR domain, the next three most commonly found domains in the *P. peruviana* transcriptome are: protein kinase domain (294 cDNAs), NB-ARC domain (190 cDNAs) and WD40 domain (123 cDNAs). Protein kinases are one of the largest protein families in plants, involved in a wide variety of physiological processes [35], like calcium-dependent protein kinases and MAP kinases which are involved in the recognition of elicitors or pathogens and the subsequent activation of defense response in plants [36]. The NB-ARC domain is a nucleotide-binding motif shared by plant resistant gene products involved in regulated cell death [37,38]. The WD40 domain, whose common function is coordinating multi-protein complex assemblies, is found in a large number of eukaryotic proteins that cover a wide variety of functions including adaptor and regulatory modules in

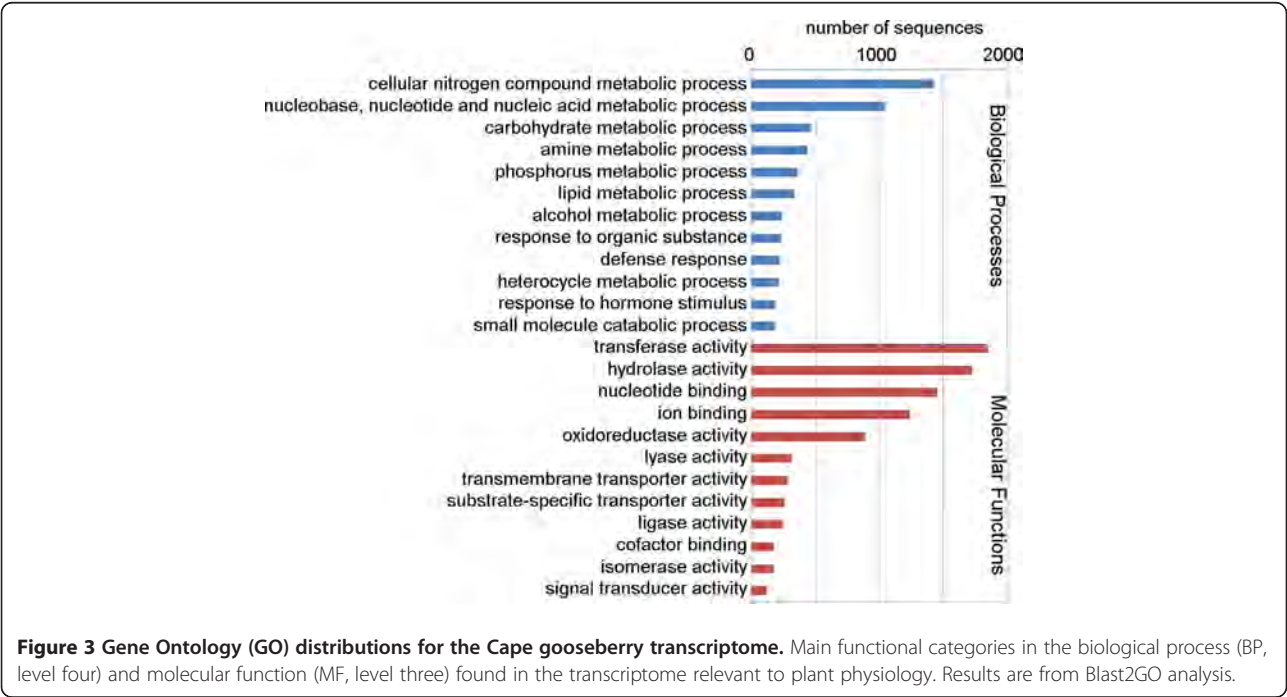
Table 3 Main metabolic pathways associated to *P. peruviana* transcripts

KEGG* metabolic pathways	Number of transcripts
General metabolic pathways	301
Purine metabolism	193
Thiamine metabolism	187
Biosynthesis of secondary metabolites	84
Biosynthesis of phenylpropanoids	53
Drug metabolism - other enzymes	48
Oxidative phosphorylation	44
Tropane, piperidine and pyridine alkaloid biosynthesis	42
Phenylalanine metabolism	25
Biosynthesis of plant hormones	14
Biosynthesis of alkaloids derived from terpenoid/polyketide	14
Biosynthesis of terpenoids and steroids	14
Other pathways	236

* KEGG: Kyoto Encyclopedia of Genes and Genomes [29].

signal transduction, pre-mRNA processing and cytoskeleton assembly [39,40]. Additionally, the WD40 domain is critically involved in the ubiquitin proteasome pathway which regulates photomorphogenesis, flowering and abiotic stress response in plants [41].

Other frequently found domains include: RNA recognition motif (115 cDNAs), RING-finger domain (96 cDNAs), Leucine rich repeat N-terminal domain (89 cDNAs), tyrosine kinase catalytic domain (84 cDNAs), all of



which are commonly found in eukaryotic cells and involved in a broad range of biological processes. The data is summarized in Table 4.

Out of the 110,921 singletons, there are 9,909 of them (length >200 bp) where GO term(s) were assigned to the sequence through Blast2GO (see Materials) or where a significant similarity to a well-characterized protein domain from NCBI CDD was found. We deposited the 9,909 singletons described above, in addition to the 24,024 assembled isotigs in the NCBI's TSA (Transcriptome Shotgun Assembly) sequence database, which is available at GenBank (accessions JO124085-JO157957). Those sequences with their functional annotations, including GO terms and domain similarity related description, are also provided as Additional file 1: 'Cape gooseberry cDNAs'.

In silico SSR marker identification

The presence of Simple Sequence Repeats (SSRs) in the *P. peruviana* transcriptome was identified *in silico* using Phobos [42]. A total of 5,971 SSR loci were found in the Cape gooseberry cDNAs, where imperfect motifs were the most abundant (5,568), in contrast to 403 loci representing perfect motifs (Table 5). Microsatellites were searched in cDNAs avoiding redundant results in isotigs, considering that searching in the alternative transcripts could lead us to predict the same SSRs in different isotigs corresponding to the same isogroup. Trinucleotide (1,068) and hexanucleotide (3,036) motifs were the most commonly found repetitions in the *P. peruviana* leaf transcriptome, accounting for 68% of the SSRs identified, in contrast to other plant studies where tri- and dinucleotides were the most commonly found repeat units [18,43,44].

We recently reported the first set of microsatellite markers developed for *P. peruviana* and related species [45] where the large majority of SSR loci was found in

Table 5 SSRs identified in *P. peruviana* cDNAs

Motif	PERFECT	IMPERFECT
Dinucleotide	69	275
Trinucleotide	150	918
Tetranucleotide	28	465
Pentanucleotide	22	1,008
Hexanucleotide	134	2,902
TOTAL	403	5,568

untranslated regions (UTRs) of transcripts with similarity to known proteins in public databases, leading to the identification of two novel polymorphic SSRs related to proteins involved in pathogen defense response. SSRs prioritization for plant breeding programs can be done via functional annotation of cDNAs associated with predicted SSRs and Gene Ontology annotations like ones involved in plant defense. Here we used and updated functional annotation of the transcriptome and the entire collection of assembled transcripts to report ten novel predictions for cDNA-derived SSRs in Cape gooseberry. These SSRs are associated with proteins with gene ontology annotations involved in plant defense to biotic stress such as defense response to fungus, programmed cell death, callose deposition in cell wall during defense response, plant hypersensitive response, and jasmonic acid, ethylene and salicylic acid hormones (Additional file 2: 'Functional annotation of ten *Physalis peruviana* SSRs markers related to plant defense'). The SSRs obtained in this study are the raw materials for future studies in genetic variation among *Physalis* populations, which can be used for: construction of genetic maps, quantitative trait loci (QTL) identification in this species and plant breeding programs focused on phytosanitary Cape gooseberry problems.

Gene model prediction in *P. Peruviana*

The genome of *P. peruviana* has not been sequenced yet, nevertheless it is possible to generate gene model predictions using the *P. peruviana* transcriptome and the genomes of *Solanum lycopersicum* (tomato) or *Solanum tuberosum* (potato), which are the two closest related species that have genome sequence available [46,47]. The cDNA to genomic DNA alignments were generated using the Splign software package [48] as described in Methods. All the assembled transcripts through the previous steps including 23,964 isotigs and 9,909 singletons, were mapped to the *S. lycopersicum* genome, resulting in 12,436 (36.7%) aligned cDNAs, representing 8,801 gene loci and 9,454 transcript models. On the other hand, 14,515 (42.9%) *P. peruviana* cDNAs were mapped to the *S. tuberosum* genome, representing 10,166 gene loci and 10,992 transcript models, as summarized in Table 6. Splign requires the consensus intron sequences (GT/AG

Table 4 Protein domains identified in *P. peruviana* transcriptome

CDD* Identifier	Domain name	Number of cDNAs
193328, 188079	Pentatricopeptide repeat domain (PPR motif)	350
173623, 189373	Protein kinase domain	294
144508	NB-ARC domain	190
29257	WD40 domain	123
128654	RNA recognition motif	115
29102	RING-finger domain	96
191981	Leucine rich repeat N-terminal domain	89
128515	Tyrosine kinase catalytic domain	84
	Others	11633

*CDD: Conserved Domain Database.

Table 6 Cape gooseberry gene model prediction overview from alignments to the tomato and potato genomes

	Aligned cDNAs	Gene loci	Transcript models
<i>S. lycopersicum</i> genome	12,436	8,801	9,454
<i>S. tuberosum</i> genome	14,515	10,166	10,992

or GC/AG) at the splice sites; therefore strand orientation for the multiple-exon alignments from the Splign output can be decided by the 4-nucleotide sequences at the two intron(s) ends. At the moment, no strand orientation is assigned to single exon transcripts, as the query sequence gets aligned to a continuous region in the genome, unless there is strong polyadenylation signal.

The majority of aligned exons have an identity with the genomic sequence ranging from 70% to 95%, with an average identity of 87.6%. Figure 4 shows a number of features of the gene models from alignments of *P. peruviana* to *S. tuberosum* genome. Most of gene models contain less than 20 exons. The longest one has 51 exons. The average length of the aligned exons is 228 base pairs and that of the intron is 1,287 base pairs. The intron-exon boundaries as predicted by cDNA to genome alignments are highly conserved when both the *S. tuberosum* and *S. lycopersicum* genomes are used to align the *P. peruviana* transcriptome. We have generated General Feature Format (GFF) files for the gene models (Additional file 3: 'Cape gooseberry gene model predictions using the tomato genome' and Additional file 4: 'Cape gooseberry gene model predictions using the potato genome').

Further examination of the gene models revealed that there are 11,949 *P. peruviana* cDNAs mapped to both *S. lycopersicum* and *S. tuberosum* genome as shown in Figure 5 (panel A-a), among which 9,795 cDNAs have multiple-exon gene models on both genomes (panel A-b). 9,436 (96%) multiple-exon cDNAs have at least one intron occurring at exactly the same position on the cDNA when aligned to the *S. lycopersicum* and *S. tuberosum* genomes (panel A-c). Furthermore, there are 6,358 cDNAs having exactly the same set of intron positions on the cDNA when mapped to the two genomes (Figure 5 panel A-d). However, for those intron positions that have the same coordinates in the cDNA when mapped to the two genomes, the length and thus the sequence of corresponding introns in the gene models from the two genomes have large variances, as revealed in Figure 5 B.

Intron length variation is exemplified in Figure 5C, where a *P. peruviana* cDNA (ID Php00a06743.16696) was mapped to both the *S. lycopersicum* and *S. tuberosum* genomes, resulting in two identical sets of exons, but different sets of intron lengths (a, b). There is also a number of *S. lycopersicum* and *S. tuberosum* cDNAs that have the same predicted gene model in their own genome, respectively (all the cDNAs are aligned by Splign). Figure 5C (c) shows the nucleotide sequences around the first intron site of the 3 cDNAs from *P. peruviana*, *S. tuberosum* and *S. lycopersicum*. Primers targeting conserved flanking exonic regions as indicated can be used to amplify intronic fragments from all three species, *P. peruviana*, *S. lycopersicum* and *S. tuberosum*.

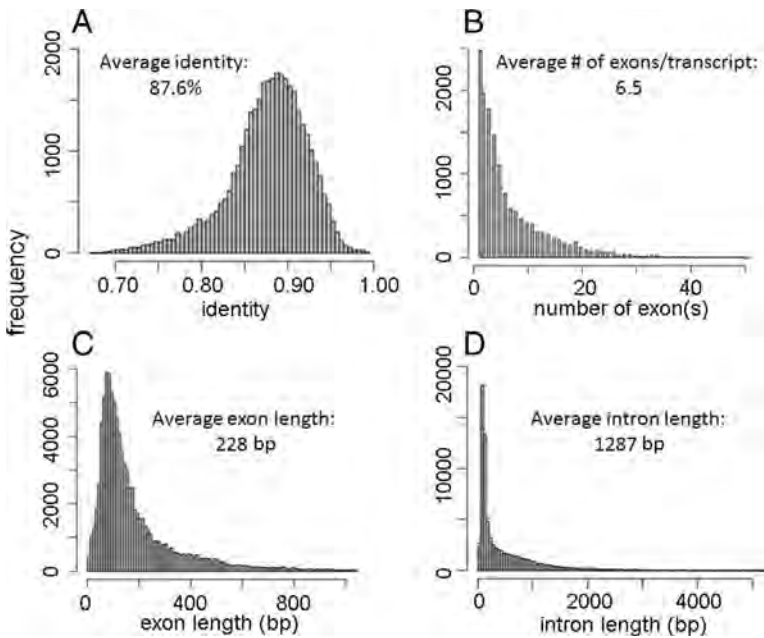


Figure 4 Predicted gene models features for alignments of *P. peruviana* to the *S. tuberosum* genome. A) Distribution of exon identities. **B)** Distribution of number of exons per transcript. **C)** Distribution of exon lengths. **D)** Distribution of intron lengths.

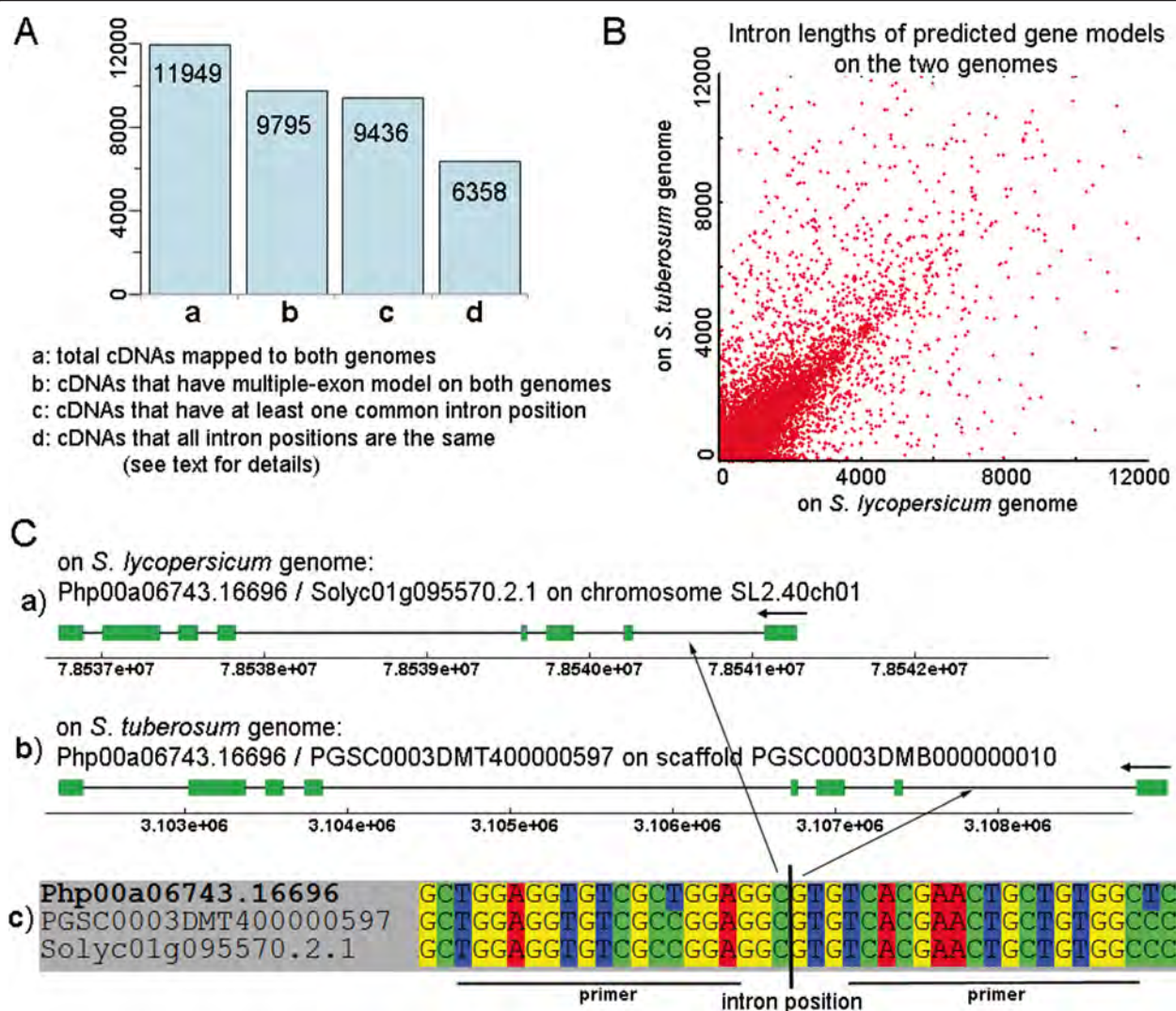


Figure 5 Gene model prediction in *P. peruviana*. **A**) The number of *P. peruviana* cDNAs that (a) can be mapped to both *S. lycopersicum* and *S. tuberosum* genomes; (b) have multiple-exon models on both genomes; (c) have at least one common intron position (on the cDNA); (d) all introns positions are the same in gene models from the two genomes. **B**) For those “common” intron points, the intron lengths in the predicted gene models on the two genomes have big variation. **C**) A typical example: (a) *P. peruviana* cDNA Php00a06743.16696 and *S. lycopersicum* cDNA Solyc01g095570.2.1 have identical gene models on *S. lycopersicum* chromosome SL2.40ch01. (b) The same *P. peruviana* cDNA and *S. tuberosum* cDNA PGSC0003DMT400000597 have identical gene models on *S. tuberosum* superscaffold PGSC0003DMB000000010. The exon sets of the *P. peruviana* cDNA in panel A and panel B are identical but the two intron sets have remarkable differences. (c) Nucleotide sequences at the first intron junction. Primers can be designed at the indicated positions to amplify intron regions from the Solanaceae.

The conserved orthologous set (COS) markers are sets of genes conserved throughout evolution in both sequence and copy number [49,50] that have been used extensively in comparative genomic and phylogenetic studies in *Solanaceae*. The COS marker strategy involves design of universal exonic primers among closely related species based on ortholog identification and multiple sequence alignment to amplify intronic/exonic regions. In the present study, we present another convenient approach to find universal exon regions - gene model prediction by Splign using two or more related genomes to define

common models. Given the fact that the *P. peruviana* genome is not available yet, and genomes of both *S. lycopersicum* and *S. tuberosum* are only in their initial versions, gene model predictions would be particularly valuable in obtaining specific intronic regions for marker and SNP discovery in non-model species, as well as for comparative genomic and phylogenetic studies.

We also aligned the *S. lycopersicum* transcriptome to its own genome (data from http://solgenomics.net/organism/solanum_lycopersicum/genome) and also to the *S. tuberosum* genome [48] using Splign. We mapped

34,704 from a total of 34,727 *S. lycopersicum* cDNAs to its own genome with 100% identity (data not shown). However, only 28,366 (81.7%) *S. lycopersicum* cDNAs can be mapped to *S. tuberosum* genome with an average identity of 90%. In the Cape gooseberry case only 42.9% *P. peruviana* cDNAs get mapped to the *S. tuberosum* genome, suggesting that *P. peruviana* is evolutionarily more distant from *S. lycopersicum* and *S. tuberosum* than the two species from each other. We then conducted further analysis to estimate the phylogenetic location of *P. peruviana* in the Solanaceae family.

Experimental validation of intron positions

The experimental validation of predicted exon/intron boundaries in the assembled cDNAs was carried out in a small sample of cDNAs, which are putative homologous of plant disease resistance genes and can be mapped to both the potato and tomato genomes. For each of these cDNAs, a pair of COSII primers was designed to span one putative intron (based on the computational predicted gene model) for PCR amplification of the genomic DNA. The information of the primers used is summarized in Table 7. All the amplified PCR products had the expected length and then were sequenced using conventional Sanger sequencing. Comparison of the amplified genomic fragments to their corresponding cDNAs revealed that all the eight samples we sequenced indeed showed the exon/intron boundaries consistent with the gene models predicted by Splign. Three of them had the experimentally identified intron positions exactly the same as the predicted. In the other five samples, the predicted intron positions are a few base pairs (1–6 bp) away from the experimentally identified sites. The results are shown in Table 8.

Phylogenetic relationship of *P. Peruviana* with other solanaceae species

We found five putative orthologs among *P. peruviana*, *S. lycopersicum*, *S. tuberosum*, *Capsicum spp* (pepper), *S. melongena* (eggplant) and *Petunia spp*. The proteins are: xyloglucanase inhibitor containing pepsin_retropepsin superfamily domain, mitochondrial catalytic protein containing PP2Cc superfamily domain, mitochondrial small ribosomal subunit protein containing RPS2 superfamily domain, phosphate transporter and a functionally unknown protein.

To obtain the best accuracy of the phylogenetic tree to be built, we compared the five putative orthologous proteins to the NCBI's plant RefSeq protein database. There are seven other species that have BLAST hits with an expect value < 1e-5 to all the five orthologs from the previous steps. These seven species are: *Arabidopsis thaliana*, *Populus trichocarpa* (black cottonwood), *Ricinus communis* (castor bean), *Vitis vinifera* (grape), *Oryza sativa* (rice), *Zea mays* (corn) and *Sorghum bicolor*, none of which belongs to the Solanaceae family. The phylogenetic tree was constructed between thirteen plants using the software Phym1 [51] and MEGA [52]. The tree generated has good bootstrapping support at all of the branch points except for the position of *V. vinifera*. We removed the *V. vinifera* sequence and constructed the tree presented in Figure 6.

The phylogenetic relationship among *S. lycopersicum*, *S. tuberosum*, *Capsicum spp*, *S. melongena* and *Petunia spp* is consistent with a previous study by Wang Y et al. [53], in which the tree was constructed based on an unduplicated conserved syntenic segment in the genomes of the five plants. Our results showed that *P. peruviana* branched out before the divergence of the other five Solanaceae family

Table 7 Primers used for experimental validation of intron positions

COSII Marker	Primer sequences (5' – 3')	<i>P. peruviana</i> unique identifier	Primer position in the cDNA
C2_At3g07100	ACGAACGATGTGCTGCTGGATATAC AGACCCTGGGGATCTAAGCTCTCTG	Php00a01046.06900	F-1156/1178 R-997/1022
C2_At 2 g35920	TGCTTGCAACCAACATAGCTGAG AAGCTCTGTAGTGGTGTTCGAAG	Php00a05845.15798	F-424/446 R-172/195
C2_At 3 g06580	TGCTCAACTCACATGTGAGTGTGAAAG AGCAAACCCAGATTTTGCCATAAC	Php00a06435.16388	F-715/740 R-784/806
C2_At 5 g41480	TATTCGTGCTGGTCTGGAGAGTGTC ATGATCCTTGTCATTCGCCATAGC	Php00a02812.11168	F-836/859 R-646/669
C2_At 5 g27620	ATCTACAATGGTCCGTGATGGAAC TTCCTCTGCCTTGCAAGCTGC	Php00a03329.12202	F-776/799 R-720/740
C2_At 3 g04870	ACGCGTGCTAGTATCCAGAGG TGACATGGCAAAGCCCACTAACATAC	Php00a02563.10671	F-1226/1246 R-954/977
C2_At 5 g60160	ACACAATGCTAATCAACGTTATGC TCATCCACCGCGCACATTTC	Php00a01985.09526	F-278/297 R-482/505

Table 8 Comparison of validated exon/intron boundaries between PCR results and the predicted gene models

<i>P. peruviana</i> unique identifier	Position of boundary of the 2 adjacent exons at the mRNA		Match predicted gene model?
	PCR results	Predicted model	
Php00a01046.06900	1077/1078	1077/1078	Precisely
Php00a05845.15798	297/298	294/295	3 bp shift
Php00a02208.09960	653/654	647/648	6 bp shift
Php00a06435.16388	774/775	774/775	Precisely
Php00a02812.11168	751/752	749/750	2 bp shift
Php00a03329.12202	756/757	756/757	Precisely
Php00a02563.10671	1040/1041	1037/1038	3 bp shift
Php00a01985.09526	416/417	417/418	1 bp shift

members. Details of the phylogenetic analysis are summarized in Additional file 5: ‘Phylogenetic analysis workflow’.

Conclusions

This report constitutes the first genomic resource for the *Physalis* genus providing a large collection of assembled and functionally annotated cDNAs. The *Physalis* genus is part of the Solanaceae family, whose members are important sources of food, spice and medicine. However, genomic data for other members of the *Physalis* genus is limited. Therefore, this resource will enhance comparative studies within the family and the transcriptome will serve as a starting point for gene discovery in *Physalis* and for future annotations of the *Physalis peruviana* genome sequence. A number of the genes identified in this study provide candidates for resistance genes against viruses, fungal or bacterial pathogens. Additionally, this study is a potentially invaluable resource for mapping and marker-assisted breeding in *Physalis peruviana* and closely related species like *Physalis philadelphica*,

commonly known as tomatillo, which are food staples in Central American countries.

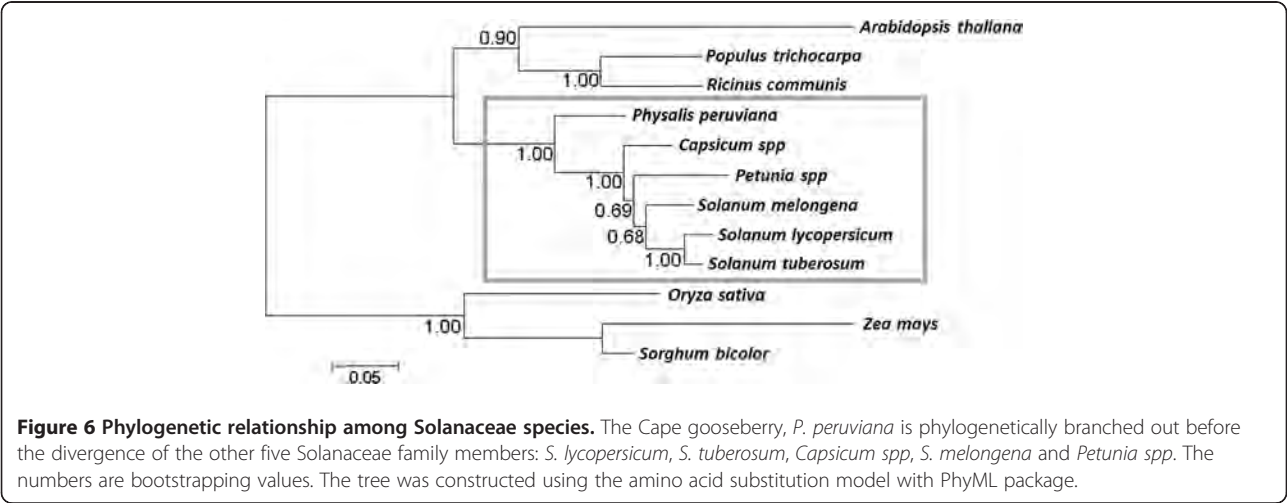
Methods

cDNA synthesis and cDNA library normalization

Fresh leaf tissue from the Cape gooseberry *Physalis peruviana* Colombian ecotype plants from the Colombian *in vitro* germplasm bank (accession number 09U216-6) at the Corporacion Colombiana de Investigacion Agropecuaria (CORPOICA) were processed and flash frozen in liquid nitrogen. Tissues were immediately sent to Bio S&T Inc. (Montreal, QC, Canada) where RNA extraction, cDNA synthesis and normalization were performed. Briefly, RNA was extracted using a modified TRIzol method (Invitrogen, USA). cDNA synthesis was carried out using 16 µg total RNA by a modified SMART™ cDNA synthesis method and then were normalized by a modified normalization method [54,55] where full-length cDNA was synthesized with two set of primers for driver and tester cDNA. Single-stranded cDNA was used for hybridization instead of double-stranded cDNA. Excess amounts of sense-stranded cDNA hybridized with anti-sense-stranded cDNA. After hybridization, duplex DNA was removed by hydroxyapatite chromatography. Normalized tester cDNA was re-amplified and purified with tester specific primer L4N by failsafe™ PCR (Epicentre Biotechnologies, USA), while driver cDNA was unable to amplify using L4N primer. Size fractionation of re-amplified cDNA was done in a 1% agarose gel. Greater than 0.5 kb cDNA fragments were purified by electroelution and after determining the concentrations, purified cDNAs were precipitated and stored in 80% EtOH at -80°C.

cDNA sequencing and assembly

The normalized cDNA library was prepared for sequencing at Emory Genomics Center (Atlanta, GA, USA).



Approximately 5 µg of purified cDNA was sheared into small fragments via Covaris E210 Acoustic Focusing Instrument and sequenced in three-fourths 454 plate run on a 454 GS-FLX Titanium platform (Roche). The SFF files containing raw sequences and quality scores were submitted to the NCBI Sequence Read Archive (accession number SRP005904).

SeqClean [19,56] was used before and after the assembly, for automated trimming and validation of the raw read files and the assembled file. SeqClean was launched with a minimum and maximum length cut-off of 50pb and 600pb. We used the Newbler software, *GS de novo Assembler* (Roche, version 2.5.3) with default parameters, to assemble reads into contigs, then further into isotigs. Isotigs within an isogroup represent putative alternatively spliced transcripts of a gene. Reads that cannot be connected with any others were defined as singletons.

Functional annotation

After assembly, a local BLASTX [22,23] was used to compare the assembled isotigs and singletons against the UniProtKB/Swiss-Prot database (released on April-2011) using an expect value threshold of $1e-5$. The remaining cDNAs that did not get hits from UniProtKB/Swiss-Prot were compared against the NCBI RefSeq database (Release 47). The BLASTX output (XML format) was subjected to Blast2GO [24] for Gene Ontology (GO) analysis. Blast2GO retrieves the most significant GO terms associated with the obtained hits to the query sequence. When possible, Blast2GO also provides Enzyme commission (EC) numbers and the metabolic pathways they participate. We also compared all the *P. peruviana* cDNAs against the NCBI CDD database [30] using an expect value threshold of $1e-5$ and selected all the hits where the aligned length is more than 2/3 of the targeted CDD length for domain identification.

SSR identification

Phobos (version 3.3.11) (http://www.rub.de/spezzoo/cm/cm_phobos.htm) was used to identify microsatellites (SSRs) in the publicly available collection of assembled transcripts and singletons [GenBank: JO124085-JO157957]. Perfect and imperfect searches were performed using default parameters.

Gene model prediction

Gene model prediction was carried out using the software package Splign [48], which has been proven to be able to accurately compute cDNA-to-Genome alignment with high efficiency. At the heart of the program is a compartmentalization algorithm which identifies possible gene duplications, and a refined alignment algorithm recognizing introns and splice signals. The complete genome of *P. peruviana* is not available yet. The two closest relatives of *P. peruviana* that have genomic sequences available are

S. lycopersicum (tomato) (data from http://solgenomics.net/organism/solanum_lycopersicum/genome; ITAG Release 2.3) and *S. tuberosum* (potato; PGSC_DM_v3.4) [47]. Therefore we used the draft genomes of potato and tomato as the reference genome to map the assembled *P. peruviana* leaf transcriptome.

Phylogenetic analysis

We selected orthologous proteins using the all-to-all alignment and mutual best hits selection strategy [57]. Pairwise alignments were performed using BLASTP (expect value $< 1e-5$) using the RefSeq proteins from *S. lycopersicum*, *S. tuberosum*, *Capsicum spp*, *S. melongena* and *Petunia spp*. At the time of analysis, the numbers of RefSeq proteins from the five species were: 4,788 for *S. tuberosum*; 6,008 for *S. lycopersicum*; 263 for *S. melongena*; 1,701 for *Capsicum spp* and 1,226 for *Petunia spp*. Fifteen putative orthologous proteins were found, which are present in all five species. Next, we aligned the assembled *P. peruviana* isotigs using BLASTX (expect value $< 1e-5$) against the database made of the fifteen orthologous groups obtained from the previous step (altogether 75 proteins). We identified eleven orthologous groups of proteins from all the five plants with hit(s) from the *P. peruviana* transcriptome. The best hit was chosen when multiple *P. peruviana* proteins hit a given group. We manually examined the alignments in eleven clusters and removed those with large length variation (the longest one is $>20\%$ of the shortest one) and susceptible similarities ($< 65\%$). Thereafter we ended up with five orthologous groups among the six species.

To obtain higher accuracy phylogenetic tree, we further compared the five orthologous groups against the entire plant RefSeq protein database using BLASTP. There are altogether seven more plants that have significant hit(s) (expect value $< 1e-5$) for all the five orthologous groups. To this step we have thirteen plants for the five orthologous groups. We concatenated the five proteins in each species (in the same order) and aligned them using the program MUSCLE [58]. The alignment results were manually refined and subjected to Phym1 [51] and MEGA version 5 [52] for phylogenetic tree construction. Bootstrapping was repeated 1,000 times. Both programs produced the same results.

Additional files

Additional file 1: Cape gooseberry cDNAs. The annotated FASTA sequences of the assembled transcriptome, including singletons.

Additional file 2: Functional annotation of ten *Physalis peruviana* SSRs markers related to plant defense.

Additional file 3: Cape gooseberry gene model predictions using the tomato genome. Gene model predictions using Splign from cDNA to genome alignments to the tomato genome.

Additional file 4: Cape gooseberry gene model predictions using the potato genome. Gene model predictions using Splign from cDNA to genome alignments to the potato genome.

Additional file 5: Phylogenetic analysis workflow.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

Support for this research was provided by a grant from the Colombian Ministry of Agriculture Contract Nos. 054/08072-2008 L4787-3281 to Luz Stella Barrero and 054/08190-2008 L7922-3322 to Victor Manuel Núñez Zarantes. Gina Garzón-Martínez was supported by a Colciencias "Joven Investigador" Fellowship during 2010. Leonardo Mariño-Ramírez expresses his deepest gratitude to his friend and colleague Dr. Alba Marina Cotes Prado for all the advice and support she gave him to conduct this project. This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and National Center for Biotechnology Information.

Author details

¹Plant Molecular Genetics Laboratory, Center of Biotechnology and Bioindustry (CBB), Colombian Corporation for Agricultural Research (CORPOICA), Bogotá, Colombia. ²Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, United States of America, Bethesda, MD, USA. ³PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia.

Authors' contributions

GG-M acquired the sequencing data, performed the genome assembly, the functional annotation, SSR marker identification and drafted the manuscript. ZIZ participated in the genome assembly, performed functional annotation, gene model prediction, phylogenetic analysis and drafted the manuscript. DL participated in the design of the study and contributed to the phylogenetic analysis. LSB conceived of the study, and participated in its design and coordination and helped to draft the manuscript. LM-R conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Received: 23 November 2011 Accepted: 25 April 2012

Published: 25 April 2012

References

- Knapp S: Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. *J Exp Bot* 2002, **53**(377):2001.
- Ramadan MF: Bioactive phytochemicals, nutritional value, and functional properties of cape gooseberry (*Physalis peruviana*): An overview. *Food Res Int* 2011, **44**(7):1830-1836.
- Wu S-J, Chang S-P, Lin D-L, Wang S-S, Hou F-F, Ng L-T: Supercritical carbon dioxide extract of *Physalis peruviana* induced cell cycle arrest and apoptosis in human lung cancer H661 cells. *Food and Chemical Toxicology* 2009, **47**(6):1132-1138.
- Yen C, Chiu C, Chang F, Chen J, Hwang C, Hseu Y, Yang H, Lee A, Tsai M, Guo Z: 4-Hydroxywithanolide E from *Physalis peruviana* (golden berry) inhibits growth of human lung cancer cells through DNA damage, apoptosis and G2/M arrest. *BMC cancer* 2010, **10**(1):46.
- He C, Saedler H: Heterotopic expression of MPF2 is the key to the evolution of the Chinese lantern of *Physalis*, a morphological novelty in Solanaceae. *Proc Natl Acad Sci U S A* 2005, **102**(16):5779-5784.
- Hu JY, Saedler H: Evolution of the inflated calyx syndrome in Solanaceae. *Mol Biol Evol* 2007, **24**(11):2443-2453.
- Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W, Shi L: Next-generation sequencing and its applications in molecular diagnostics. *Expert Rev Mol Diagn* 2011, **11**(3):333-343.
- Imelfort M, Edwards D: De novo sequencing of plant genomes using second-generation technologies. *Briefings in Bioinformatics* 2009, **10**(6):609-618.
- Blanca J, Canizares J, Roig C, Zarsolo P, Nuez F, Pico B: Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). *BMC Genomics* 2011, **12**:104.
- Li Y, Luo H, Sun C, Song J, Sun Y, Wu Q, Wang N, Yao H, Steinmetz A, Chen S: EST analysis reveals putative genes involved in glycyrrhizin biosynthesis. *BMC Genomics* 2010, **11**(1):268.
- Novaes E, Drost D, Farmerie W, Pappas G, Grattapaglia D, Sederoff R, Kirst M: High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 2008, **9**(1):312.
- Parchman T, Geist K, Grahnen J, Benkman C, Buerkle C: Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* 2010, **11**(1):180.
- Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui E, Chen S: De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. *BMC Genomics* 2010, **11**(1):262.
- Alagna F, D'Agostino N, Torchia L, Servili M, Rao R, Pietrella M, Giuliano G, Chiusano ML, Baldoni L, Perrotta G: Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics* 2009, **10**:399.
- Wang W, Wang Y, Zhang Q, Qi Y, Guo D: Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics* 2009, **10**(1):465.
- Rismani-Yazdi H, Haznedaroglu BZ, Bibby K, Peccia J: Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: pathway description and gene discovery for production of next-generation biofuels. *BMC Genomics* 2011, **12**:148.
- Cheung F, Haas BJ, Goldberg SM, May GD, Xiao Y, Town CD: Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics* 2006, **7**:272.
- Zeng S, Xiao G, Guo J, Fei Z, Xu Y, Roe BA, Wang Y: Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics* 2010, **11**:94.
- SeqClean [http://compbio.dfci.harvard.edu/tgi/software/]
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005, **437**(7057):376-380.
- Kumar S, Blaxter ML: Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics* 2010, **11**:571.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: BLAST+: architecture and applications. *BMC Bioinforma* 2009, **10**:421.
- Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, **21**(18):3674-3676.
- Aharoni A, Galili G: Metabolic engineering of the plant primary-secondary metabolism interface. *Curr Opin Biotechnol* 2011, **22**(2):239-244.
- Ahn IP, Kim S, Lee YH: Vitamin B1 functions as an activator of plant disease resistance. *Plant Physiol* 2005, **138**(3):1505-1515.
- Goyer A: Thiamine in plants: aspects of its metabolism and functions. *Phytochemistry* 2010, **71**(14-15):1615-1624.
- Korkina LG: Phenylpropanoids as naturally occurring antioxidants: from plant defense to human health. *Cell Mol Biol (Noisy-le-grand)* 2007, **53**(1):15-25.
- Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2011.
- Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, et al: CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 2011, **39**(Database issue):D225-D229.
- Shikanai T, Okuda K: In Vitro RNA-Binding Assay for Studying Trans-Factors for RNA Editing in Chloroplasts. *Methods Mol Biol* 2011, **774**:199-208.
- Zehrmann A, Verbitskiy D, Hartel B, Brennicke A, Takenaka M: PPR proteins network as site-specific RNA editing factors in plant organelles. *RNA Biol* 2011, **8**(1):67-70.
- Takenaka M, Verbitskiy D, Zehrmann A, Brennicke A: Reverse Genetic Screening Identifies Five E-class PPR Proteins Involved in RNA Editing in Mitochondria of *Arabidopsis thaliana*. *J Biol Chem* 2010, **285**(35):27122-27129.
- Fujii S, Small I: The evolution of RNA editing and pentatricopeptide repeat genes. *New Phytol* 2011, **191**(1):37-47.
- Chevalier D, Walker JC: Functional genomics of protein kinases in plants. *Brief Funct Genomic Proteomic* 2005, **3**(4):362-371.
- Romeis T: Protein kinases in the plant defence response. *Current opinion in plant biology* 2001, **4**(5):407-414.
- Liu J, Liu X, Dai L, Wang G: Recent progress in elucidating the structure, function and evolution of disease resistance genes in plants. *J Genet Genomics* 2007, **34**(9):765-776.

38. van der Biezen EA, Jones JD: The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr Biol* 1998, **8**(7):R226–R227.
39. Stirnimann CU, Petsalaki E, Russell RB, Muller CW: WD40 proteins propel cellular networks. *Trends Biochem Sci* 2010, **35**(10):565–574.
40. Xu C, Min J: Structure and function of WD40 domain proteins. *Protein Cell* 2011, **2**(3):202–214.
41. Biedermann S, Hellmann H: WD40 and CUL4-based E3 ligases: lubricating all aspects of life. *Trends Plant Sci* 2011, **16**(1):38–46.
42. Phobos 3.3.11 [http://www.rub.de/spezzoo/cm/cm_phobos.htm]
43. Argout X, Fouet O, Wincker P, Gramacho K, Legavre T, Sabau X, Risterucci AM, Da Silva C, Cascardo J, Allegre M: Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions. *BMC Genomics* 2008, **9**(1):512.
44. Luro FL, Costantino G, Terol J, Argout X, Allario T, Wincker P, Talon M, Ollitrault P, Morillon R: Transferability of the EST-SSRs developed on Nules clementine (*Citrus clementina* Hort ex Tan) to other Citrus species and their effectiveness for genetic mapping. *BMC Genomics* 2008, **9**(1):287.
45. Simbaqueba J, Sanchez P, Sanchez E, Nunez Zaranter VM, Chacon MI, Barrero LS, Marino-Ramirez L: Development and Characterization of Microsatellite Markers for the Cape Gooseberry *Physalis peruviana*. *PLoS One* 2011, **6**(10):e26719.
46. Mueller LA, Tanksley SD, Giovannoni JJ, van Eck J, Stack S, Choi D, Kim BD, Chen M, Cheng Z, Li C, *et al*: The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL). *Comp Funct Genomics* 2005, **6**(3):153–158.
47. Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, *et al*: Genome sequence and analysis of the tuber crop potato. *Nature* 2011, **475**(7355):189–195.
48. Kapustin Y, Souvorov A, Tatusova T, Lipman D: Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol Direct* 2008, **3**:20.
49. Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD: Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* 2002, **14**(7):1457–1467.
50. Wu F, Mueller LA, Crouzillat D, Petiard V, Tanksley SD: Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSIL) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics* 2006, **174**(3):1407–1420.
51. Boussau B, Gueguen L, Gouy M: A mixture model and a hidden markov model to simultaneously detect recombination breakpoints and reconstruct phylogenies. *Evol Bioinform Online* 2009, **5**:67–79.
52. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* 2011, **10**:2731–2739.
53. Wang Y, Diehl A, Wu F, Vrebalov J, Giovannoni J, Siepel A, Tanksley SD: Sequencing and comparative analysis of a conserved syntenic segment in the Solanaceae. *Genetics* 2008, **180**(1):391–408.
54. Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A: Construction and characterization of a normalized cDNA library. *Proc Natl Acad Sci U S A* 1994, **91**(20):9228–9232.
55. Patanjali SR, Parimoo S, Weissman SM: Construction of a uniform-abundance (normalized) cDNA library. *Proc Natl Acad Sci U S A* 1991, **88**(5):1943–1947.
56. Chen YA, Lin CC, Wang CD, Wu HB, Hwang PI: An optimized procedure greatly improves EST vector contamination removal. *BMC Genomics* 2007, **8**:416.
57. Li L, Stoeckert CJ Jr, Roos DS: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003, **13**(9):2178–2189.
58. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004, **32**(5):1792–1797.

doi:10.1186/1471-2164-13-151

Cite this article as: Garzón-Martínez *et al*: The *Physalis peruviana* leaf transcriptome: assembly, annotation and gene model prediction. *BMC Genomics* 2012 **13**:151.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



Servicio académico como Editor de revistas internacionales

Evidencia de publicaciones de investigadores colombianos donde el Dr. Leonardo Mariño Ramírez ha servido como Editor



Identification of the *Plasmodium falciparum* rhoptry neck protein 5 (PfRON5)

Hernando Curtidor^{a,b,c,1}, Liliana C. Patiño^{a,b,1}, Gabriela Arévalo-Pinzón^{a,b},
Manuel E. Patarroyo^{a,d}, Manuel A. Patarroyo^{a,b,*}

^a Fundacion Instituto de Inmunologia de Colombia, Carrera 50 No. 26-20, Bogota, Colombia

^b Universidad del Rosario, Calle 14 No. 6-25, Bogota, Colombia

^c Universidad de la Sabana, Km. 7, Autopista Norte, Bogota, Colombia

^d Universidad Nacional de Colombia, Carrera 45 No. 26-85, Bogota, Colombia

ARTICLE INFO

Article history:

Accepted 14 December 2010

Available online 23 December 2010

Received by Leonardo Marino-Ramirez

Keywords:

Malaria

Plasmodium falciparum

Merozoite

Erythrocyte invasion

Homology

ABSTRACT

Gathering knowledge about the proteins involved in erythrocyte invasion by *Plasmodium* merozoites is the starting point for developing new strategies to control malarial disease. Many of these proteins have been studied in *Toxoplasma gondii*, where some belonging to the Moving Junction complex have been identified. This complex allows a strong interaction between host cell and parasite membranes, required for parasite invasion. In this genus, four rhoptry proteins (RON2, RON4, RON5 and RON8) and one micronemal protein (TgAMA-1) have been found as part of the complex. In *Plasmodium falciparum*, RON2 and RON4 have been characterized. In the present study, we identify PfRON5, a ~110 kDa protein which is expressed in merozoite and schizont stages of the FCB-2 strain.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Among the five parasites causing malaria in humans, *Plasmodium falciparum* is the species responsible for the highest morbidity; ~250 million cases of malaria are reported per year and a 93% of them are attributed to this parasitic species (WHO, 2009). Different strategies to eradicate this disease have been designed, such as pesticides, mosquito nets, antimalarial drugs, and different types of vaccines. However, several factors such as the global expansion of the disease (due in part to the increase of strains resistant to antimalarial drugs), the parasite's high genetic variability, the vector resistance to insecticides and the poor socioeconomic conditions of affected populations justify the search and adoption of new measures, such as the development of a fully effective vaccine (Good, 2001).

Merozoite surface proteins and some rhoptry and micronemal proteins have been considered as vaccine targets, since they are

exposed to the immune system during the invasion of red blood cells (Cowman et al., 2002). This invasion begins with the merozoite reversible binding to the erythrocyte surface, mainly mediated by the Merozoite Surface Proteins (MSPs) (Chitnis and Blackman, 2000). Subsequently, a high affinity binding known as Tight Junction (TJ) occurs between the merozoite apical end and the erythrocyte membrane. The TJ migrates from the anterior to the posterior end of the merozoite during invasion (Moving Junction) activating the actin-myosin machinery (Alexander et al., 2005; Baum et al., 2008; Straub et al., 2009). As the parasite is moving in, the parasitophorous vacuole is formed, in which the parasite will develop and replicate for the next cell generation (Kaneko, 2007).

Recently, some authors have identified the proteins present in *Toxoplasma gondii* TJ complex (another member of the Apicomplexa phylum), using immunoprecipitation techniques. This TJ complex is formed by a micronemal protein, known as the apical membrane antigen 1 (AMA-1), and four rhoptry neck proteins (TgRON2, 4, 5 and 8) (Alexander et al., 2005; Baum et al., 2008; Straub et al., 2009). Besteiro and coworkers have recently proposed a model for TJ proteins organization, where TgRON2 and TgRON5 are exported to the host cell membrane, while TgRON4 and 8 are translocated to the host cell cytoplasm. In this model, TgRON2 and TgRON5 are exposed to the host cell surface and TgRON2 acts as a specific receptor for TgAMA-1, which is located on the parasite membrane (Besteiro et al., 2009).

The presence of orthologous genes which encode for RON proteins in different Apicomplexa members suggests that the TJ complex formation is a conserved mechanism in this phylum (Proellocks et al., 2010). TgRON2 and TgRON4 homologous proteins have been

Abbreviations: AMA-1, Apical membrane antigen-1; DAPI, 4',6-Diamidino-2-phenylindole; EGTA, Ethylene glycol tetraacetic acid; ELISA, Enzyme-linked immunosorbent assay; FCA, Freund's complete adjuvant; FIA, Freund's incomplete adjuvant; FITC, Fluorescein isothiocyanate; IFA, Indirect immunofluorescence assay; MSPs, Merozoite surface proteins; PBS, Phosphate buffered saline; RBC, Red blood cell; RON, Rhoptry neck protein; RP-HPLC, Reverse phase-high performance liquid chromatography; RT-PCR, Reverse transcription polymerase chain reaction; TMs, Transmembrane domains.

* Corresponding author. Carrera 50 No. 26-20, Bogotá, Colombia. Fax: +57 1 4815269.

E-mail address: mapatarrr.fidic@gmail.com (M.A. Patarroyo).

¹ Both authors contributed equally to this work.

Identification of *Plasmodium vivax* Proteins with Potential Role in Invasion Using Sequence Redundancy Reduction and Profile Hidden Markov Models

Daniel Restrepo-Montoya^{1,2,3,4}, David Becerra^{1,2}, Juan G. Carvajal-Patiño^{1,3,4}, Alvaro Mongui^{4*}, Luis F. Niño^{1,2}, Manuel E. Patarroyo^{4,5}, Manuel A. Patarroyo^{3,4*}

1 Bioinformatics and Intelligent Systems Research Laboratory - BIOLISI, Universidad Nacional de Colombia, Bogotá D.C., Colombia, **2** Research Group on Combinatorial Algorithms - ALGOS-UN, Universidad Nacional de Colombia, Bogotá D.C., Colombia, **3** School of Medicine and Health Sciences, Universidad del Rosario, Bogotá D.C., Colombia, **4** Fundación Instituto de Inmunología de Colombia - FIDIC, Bogotá D.C., Colombia, **5** School of Medicine, Universidad Nacional de Colombia, Bogotá D.C., Colombia

Abstract

Background: This study describes a bioinformatics approach designed to identify *Plasmodium vivax* proteins potentially involved in reticulocyte invasion. Specifically, different protein training sets were built and tuned based on different biological parameters, such as experimental evidence of secretion and/or involvement in invasion-related processes. A profile-based sequence method supported by hidden Markov models (HMMs) was then used to build classifiers to search for biologically-related proteins. The transcriptional profile of the *P. vivax* intra-erythrocyte developmental cycle was then screened using these classifiers.

Results: A bioinformatics methodology for identifying potentially secreted *P. vivax* proteins was designed using sequence redundancy reduction and probabilistic profiles. This methodology led to identifying a set of 45 proteins that are potentially secreted during the *P. vivax* intra-erythrocyte development cycle and could be involved in cell invasion. Thirteen of the 45 proteins have already been described as vaccine candidates; there is experimental evidence of protein expression for 7 of the 32 remaining ones, while no previous studies of expression, function or immunology have been carried out for the additional 25.

Conclusions: The results support the idea that probabilistic techniques like profile HMMs improve similarity searches. Also, different adjustments such as sequence redundancy reduction using Pisces or Cd-Hit allowed data clustering based on rational reproducible measurements. This kind of approach for selecting proteins with specific functions is highly important for supporting large-scale analyses that could aid in the identification of genes encoding potential new target antigens for vaccine development and drug design. The present study has led to targeting 32 proteins for further testing regarding their ability to induce protective immune responses against *P. vivax* malaria.

Citation: Restrepo-Montoya D, Becerra D, Carvajal-Patiño JG, Mongui A, Niño LF, et al. (2011) Identification of *Plasmodium vivax* Proteins with Potential Role in Invasion Using Sequence Redundancy Reduction and Profile Hidden Markov Models. PLoS ONE 6(10): e25189. doi:10.1371/journal.pone.0025189

Editor: Leonardo Mariño-Ramírez, National Institutes of Health, United States of America

Received: July 8, 2011; **Accepted:** August 29, 2011; **Published:** October 3, 2011

Copyright: © 2011 Restrepo-Montoya et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors would like to extend their sincerest gratitude to Asociación de Investigación Solidaria (SADAR), Caja Navarra (Navarra, Spain) and Agencia Española de Cooperación Internacional para el Desarrollo for supporting and financing this project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mapatarr.fidic@gmail.com

‡ Current address: Corporación Corpogen, Bogotá D.C., Colombia

Introduction

Human malaria is caused by five parasite species from the genus *Plasmodium*, of which *Plasmodium falciparum* has a preferential distribution in African countries and is particularly important, since it produces most of the fatal cases. The second species in clinical importance for humans is *Plasmodium vivax* (predominantly distributed throughout Asia and America). *P. vivax* does not cause such an imminent life-threatening condition as that caused by *P. falciparum*; however, it imposes an important social and economic toll on the world's poorest countries, as reflected in the large number of disability adjusted life years (DALYs) associated with its

incidence [1]. Furthermore, several aspects still hamper the total eradication of this disease, which include (1) the gradual emergence of antimalarial drug resistance among parasite strains, as well as (2) insecticide-resistant populations of the malaria mosquito vector, and (3) the lack of an effective vaccine [2].

Progress in *P. vivax* research has been notably delayed by contrast with *P. falciparum*, partly due to the difficulty of establishing a long-term *in vitro* culture of this species given that it is restricted to invading human reticulocytes which only account for ~1–2% of circulating red blood cells. This difficulty has been reflected in the delayed release of its genome sequence [3], the transcriptional profile of its intra-erythrocyte developmental cycle



Methods paper

The evolution of the immune-type gene family *Rhamnospondin* in cnidariansJavier A. López^a, Matthew G. Fain^b, Luis F. Cadavid^{a,c,*}^a Department of Biology, Universidad Nacional de Colombia, Bogotá, Colombia^b Department of Biology, The University of New Mexico, Albuquerque, NM 87131, USA^c Institute of Genetics, Universidad Nacional de Colombia, Bogotá, Colombia

ARTICLE INFO

Article history:

Accepted 29 November 2010

Available online 9 December 2010

Received by Leonardo Marino-Ramírez

Keywords:

*Rhamnospondin**Hydractinia*

Immunity

Evolution

Pattern recognition receptors

ABSTRACT

Rhamnose-binding lectins (RBLs) in vertebrates function in immunity as pattern recognition receptors, opsonization agents, and activators of pro-inflammatory cytokines. Although they have been identified in some invertebrate taxa, their distribution, function, and evolutionary patterns in basal metazoans, remain largely unknown. A unique RBL-containing protein composed of 8 thrombospondin type 1 repeats (TSRs) and a single RBL domain has been identified in the colonial hydroid *Hydractinia symbiolongicarpus*. This *Rhamnospondin* (*Rsp*) gene was specifically and constitutively expressed in the mouth of feeding polyps. Here we report the full characterization of a second *Rsp* gene from a *H. symbiolongicarpus* BAC library. *Rsp1* and *Rsp2* were 1.1 kb apart, shared the same domain architecture and were 93% identical. Introns differed substantially in size and sequence, excepting two introns that were nearly identical, suggesting the action of inter-locus recombination. Sequencing full-length cDNAs from a wild-type individual corroborated the exon boundaries predicted from genomic DNA and showed gene polymorphism at both loci. Database searches and phylogenetic analyses showed that *Rsp* was found only in hydrozoans, indicating that it is an innovation of the cnidarian class Hydrozoa. Phylogenetic analysis of *Rsp* sequences in hydroids show a tendency of clustering paralogous genes, suggesting that they have evolved by concerted evolution.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Rhamnose-binding lectins (RBLs) are a family of L-rhamnose recognizing proteins first isolated from the eggs of the sea urchin *Anthodiaris crassispina* (Ozeki et al., 1991) and later identified in a variety of vertebrate and invertebrate animals. RBLs are found as a single carbohydrate-binding domain, as two or three tandemly repeated domains, or as part of multi-domain proteins. The carbohydrate-binding domain consists of about 95 amino acids having 8 highly conserved cysteine residues and two conserved N- and C-terminal motifs (Tateno et al., 2002a). Internal disulfide bonds create a unique α/β fold with long loops important for sugar recognition. RBLs have been well studied in fish and have proved to play an important role in immunity. In these species, RBLs are constitutively expressed by cells of the immune system such as lymphocytes, monocytes, neutrophils, gill mucous cells, intestine goblet cells, spleen cells, and thrombocytes (Hosono et al., 1999; Okamoto et al., 2005). In addition, RBL expression can be induced in peritoneal macrophages after an inflammatory reaction. RBLs recognize lipopolysaccharides (LPS) and lipoteichoic acid from Gram-

negative and Gram-positive bacteria, displaying preferential affinity for some types of LPSs, such as the ones found in *Escherichia coli* K-12 and *Shigella flexneri* 1A strains (Tateno et al., 2002b). In addition, it has been shown that RBLs bind to glycolipids and glycoproteins of the parasitic microsporidian *Glugea plecoglossi* (Watanabe et al., 2008). In a recent study, it was demonstrated that RBLs from the chum salmon (*Oncorhynchus keta*) induced the expression of the inflammatory cytokines IL-1, IL-8 and TNF- α in cell lines derived from rainbow trout (*Oncorhynchus mykiss*) peritoneal macrophages and gonadal fibroblasts (Watanabe et al., 2009). This induction was mediated by the recognition of globotriaosylceramide (Gb3), a sphingolipid located in the cell membrane lipid rafts. Moreover, RBLs displayed opsonic properties on the macrophage cell lines by interacting with Gb3 on the cell membrane. It has also been suggested that RBLs secreted by haemocytes from the ascidian *Botryllus schlosseri* can act as opsonins (Gasparini et al., 2008; Menin and Ballarin, 2008). Thus, RBLs are likely to act at various events of inflammatory reactions, from recognition of molecular patterns to opsonization and activation of pro-inflammatory cytokines.

Previously, we characterized an RBL-encoding gene in the colonial hydroid *Hydractinia symbiolongicarpus* (Cnidaria: Hydrozoa). This hydroid is a model animal system to study immune mechanisms in basal metazoans, and is found in near-shore waters of the northeastern United States, growing as a surface incrustation of gastropod shells occupied by pagurid hermit crabs. Colonies are diploblastic, dioecious, and composed by a network of polyps interconnected through endodermal canals.

Abbreviations: RBL, Rhamnose-binding lectin; TSR, Thrombospondin type-1 repeat; *Rsp*, *Rhamnospondin*.

* Corresponding author. Institute of Genetics, Universidad Nacional de Colombia, Cr 30 # 45-08, Ed. 426, Bogotá, DC, Colombia. Tel.: +57 1 3165000; fax: +57 1 3165532.

E-mail address: lfcadavid@unal.edu.co (L.F. Cadavid).



Evolutionary patterns of killer cell Ig-like receptor genes in Old World monkeys

Catalina Palacios^{a,b}, Laura C. Cuervo^a, Luis F. Cadavid^{a,b,*}

^a Department of Biology, Universidad Nacional de Colombia, Cr. 30 # 45-03, Bogotá, Colombia

^b Institute of Genetics, Universidad Nacional de Colombia, Cr. 30 # 45-03, Bogotá, Colombia

ARTICLE INFO

Article history:

Accepted 14 December 2010

Available online 24 December 2010

Received by Leonardo Marino-Ramirez

Keywords:

Killer cell Ig-Like receptor

Old World monkeys

Evolution

Vervet monkey

Olive baboon

Colobus monkey

ABSTRACT

Killer cell Ig-like receptors (KIRs) modulate the cytotoxic effects of Natural Killer cells. KIR genes are encoded in the Leucocyte Receptor Complex and are characterized by their high haplotypic diversity and polymorphism. The KIR system has been studied in only three species of Old World monkeys, the rhesus macaque, the cynomolgus macaque, and the sabaeus monkey, displaying a complexity rivaling that of hominids (human and apes). Here we analyzed bacterial artificial chromosome draft sequences spanning the KIR haplotype of three other Old World monkeys, the vervet monkey (*Chlorocebus aethiops*), the olive baboon (*Papio anubis*) and the colobus monkey (*Colobus guereza*). A total of 25 KIR gene models were identified in these species, predicted to encode receptors with 1, 2, and 3 extracellular Ig domains, all of them with long cytoplasmic domains having two putative ITIMs, although three had a positively charged residue in the transmembrane domain. Sequence and phylogenetic analyses showed that most Old World monkeys shared five classes of KIR loci: i) *KIR2DL5/3DL20* in the most centromeric region, followed by ii) the single Ig domain-encoding locus *KIR1D*, iii) the pseudogene *KIR2DP*, iv) the conserved *KIR2DL4*, and v) the highly diversified *KIR3DL/H* loci in the telomeric half of the cluster. An exception to this pattern was the KIR haplotype of the colobus monkey that lacked the *KIR1D*, *KIR2DP*, and *KIR2DL4* loci of the central region of the cluster. Thus, Old World monkeys display a broad spectrum of KIR haplotype variation that has been generated upon an ancestral haplotype architecture by gene duplication, gene deletion, and non-homologous recombination.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Killer cell Ig-like receptors (KIRs) play a key role modulating the cytotoxic response of Natural Killer (NK) cells against self-altered cells. They interact with MHC class I molecules on the surface of the target cell triggering either an activating response that results in cytotoxicity or an inhibitory signaling conducting to anergy (Moretta et al., 2003). KIRs are typically constituted by two (2D) or three (3D) extracellular immunoglobulin (Ig) domains, a stem region, a transmembrane region, and a long (L) or short (S) cytoplasmic domain. KIRs with two Ig domains are of two types, one lacking the D0 domain due to the presence of a pseudodexon 3, and the other lacking the D1 domain due to a deletion of exon 4 (Vilches and Parham, 2002). Receptors with long cytoplasmic domains contain two immunoreceptor tyrosine-based inhibitory motifs (ITIMs) and are therefore inhibitory, whereas KIRs with short cytoplasmic domains do not have ITIMs, are activating,

and possess a positively charged amino acid in their transmembrane region allowing the interaction with adaptor molecules such as DAP12/10 that contain immunoreceptor tyrosine-based activation motifs (ITAMs) (Lanier, 2003). In primates, KIRs are encoded within the Leukocyte Receptor Complex (LRC), flanked by the Leukocyte Ig-like Receptor (*LILR*) gene cluster at the centromeric end, and by the IgA Fc receptor gene (*FcAR*) at the telomeric end (Trowsdale et al., 2001). There are two groups of KIR haplotypes in humans that vary in gene content from 7 to 14 loci, although they have in common three framework genes, *KIR3DL3* in the 5' end, *KIR2DL4* in the central region, and *KIR3DL2* in the 3' end (Vilches and Parham, 2002). These KIR framework genes delimit two regions where most variation in gene content occurs (Abi-Rached and Parham, 2005). This KIR cluster organization has been relatively well conserved in the other hominids, i.e., chimpanzees (Sambrook et al., 2005), bonobos (Rajalingam et al., 2001), gorillas (Rajalingam et al., 2004), and orangutans (Guethlein et al., 2007).

KIRs have been studied in only three species of Old World monkeys (Cercopithecidae), the rhesus monkey (Hershberger et al., 2001; Sambrook et al., 2005; Blokhuis et al., 2010; Kruse et al., 2010), the cynomolgus macaque (Bimber et al., 2008; Campbell et al., 2009), and the West African sabaeus monkey (Hershberger et al., 2005). The rhesus monkey has the best-characterized KIR system among these species, including a completely sequenced KIR haplotype (Sambrook et al., 2005). The rhesus monkey KIR system displays high haplotypic

Abbreviations: KIR, killer cell Ig-like receptors; NK, Natural Killer; ITIM, immunoreceptor tyrosine-based inhibitory motif; ITAM, immunoreceptor tyrosine-based activation motif; MHC, major histocompatibility complex; BAC, bacterial artificial chromosome.

* Corresponding author. Institute of Genetics, Universidad Nacional de Colombia, Cr. 30 # 45-03, Ed. 426, Bogotá, Colombia. Fax: +57 1 3165526.

E-mail address: lfcadavid@unal.edu.co (L.F. Cadavid).



Review

Mechanisms of genetically-based resistance to malaria

Carolina López^{a,b,c}, Carolina Saravia^{a,b}, Andromeda Gomez^a, Johan Hoebeke^d, Manuel A. Patarroyo^{a,b,*}^a Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá, Colombia^b Universidad del Rosario, Carrera 24 No. 63C-69, Bogotá, Colombia^c Postgraduate Program in Microbiology, Universidad Nacional de Colombia, Carrera 45 No. 26-85, Bogotá, Colombia^d CNRS, Unité Propre de Recherche 9021, Institut de Biologie Moléculaire et Cellulaire, Laboratory of Therapeutic Chemistry and Immunology, F-67084 Strasbourg, France

ARTICLE INFO

Article history:

Accepted 13 July 2010

Available online 22 July 2010

Received by Leonardo Marino-Ramirez

Keywords:

Malaria

Hemoglobinopathy

Erythrocyte polymorphism

Natural resistance

ABSTRACT

Malaria remains one of the most prevalent parasitoses worldwide. About 350 to 500 million febrile episodes are observed yearly in African children alone and more than 1 million people die because of malaria each year. Multiple factors have hampered the effective control of this disease, some of which include the complex biology of the *Plasmodium* parasites, their high polymorphism and their increasingly high resistance to antimalarial drugs, mainly in endemic regions. The ancient interaction between malarial parasites and humans has led to the fixation in the population of several inherited alterations conferring protection against malaria. Some of the mechanisms underlying protection against this disease are described in this review for hemoglobin-inherited disorders (thalassemia, sickle-cell trait, HbC and HbE), erythrocyte polymorphisms (ovalocytosis and Duffy blood group), enzymopathies (G6PD deficiency and PK deficiency) and immunogenetic variants (HLA alleles, complement receptor 1, NOS2, tumor necrosis factor- α promoter and chromosome 5q31–q33 polymorphisms).

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Malaria is a parasitic disease transmitted by Anopheline mosquitoes and is highly widespread throughout tropical and subtropical regions. The exact magnitude of the problem still remains unknown (Carvalho et al., 2002) since this disease is most commonly found in poor countries (Olumese, 2005) having less developed health systems and control strategies (Phillips, 2001). In these areas, the high rates of morbidity and mortality can be mainly attributable to the lack of access to effective treatment (Suh et al., 2004) and to the growing parasite resistance to antimalarial drugs such as chloroquine and pyrimethamine (Smith et al., 2002). According to the World Malaria Report 2008, published by the World Health Organization and

UNICEF, 3.3 billion people living in 109 countries or territories were at risk of acquiring malaria by the end of 2006. It has been calculated that 250 million clinical episodes of malaria occur each year (mainly due to *Plasmodium falciparum* and *Plasmodium vivax* infections), of which more than 1 million people die (WHO and UNICEF, 2008).

The immune response induced in humans by infection caused by malarial parasites is complex and varies depending on the level of endemicity, epidemiological factors, genetic makeup, host age, parasite stage and parasite species. Repeated infection and continuous exposure are required to achieve clinical immunity (which reduces the risk of death from malaria and reduces the intensity of the clinical symptoms) and later anti-parasitic immunity (which directly reduces the numbers of parasites in an infected individual or inhibits parasite replication) (Mohan and Stevenson, 1998). Both innate and acquired immunity processes are invoked during the infection. Resistance involves genetically-based resistance mechanisms and cell-mediated immunological mechanisms, but also specific antibodies, which are able to reduce the severity of the symptoms and mortality are found among the main actors in the acquired immune response (Smith et al., 2002).

Innate immunity can be defined as being the host cells' ability to resist infection by the parasite, irrespective of their previous exposure to it (the review by Stevenson and Riley, 2004 gives detailed information about innate immunity against malaria). Resistance mechanisms have been described in both sporozoite entry to liver cells and erythrocyte invasion by merozoites (Yuthavong and Wilairat, 1993) (Fig. 1). Genetically-based resistance is involved in either altering erythrocyte invasion by merozoites, in lowering parasite

Abbreviations: 6PGL, 6-phosphogluconolactonase; CR1, Complement receptor-1; CSF, Granulocyte-macrophage colony-stimulating factor; DARC, Duffy antigen receptor for chemokines; DBL, Duffy binding like; DBP, Duffy binding protein; G6PD, Glucose-6-phosphate dehydrogenase; HBB, Hemoglobin beta gene; HbC, Hemoglobin C; HbE, Hemoglobin E; HbF, Fetal hemoglobin; HbH, Hemoglobin H; HbS, Hemoglobin S; HE, Hereditary elliptocytosis; HLA, Human leukocyte antigen; ICs, Immune complexes; MHC, Major histocompatibility complex; NADP, Nicotinamide adenine dinucleotide phosphate; NADPH, Reduced form of NADP; NO, Nitric oxide; NOS, NO synthase; NOS2, NO synthase 2; PBMCs, Peripheral blood mononuclear cells; PEP, Phosphoenolpyruvate; Pfb, *P. falciparum* blood infection; PfEMP-1, *P. falciparum* erythrocyte membrane protein-1; Pfl1, *P. falciparum* infection level 1; PK, Pyruvate kinase; PMM, Prior mild malaria; PSM, Prior severe malaria; SAO, Southeast Asian ovalocytosis; SCD, Sickle-cell disease; SNP, Single nucleotide polymorphism; TNF, Tumor necrosis factor.

* Corresponding author. Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá, Colombia. Fax: +57 1 4815269.

E-mail address: mapatarr.fidic@gmail.com (M.A. Patarroyo).



Characterisation and comparative analysis of MHC-DPA1 exon 2 in the owl monkey (*Aotus nancymae*)

Carlos F. Suárez M.^{a,b}, Manuel A. Patarroyo^{a,b}, Manuel E. Patarroyo^{a,c,*}

^a Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá, Colombia

^b Universidad del Rosario, Calle 63D No. 24-31, Bogotá, Colombia

^c Universidad Nacional de Colombia, Carrera 45 No. 26-85 Bogotá, Colombia

ARTICLE INFO

Article history:

Accepted 17 September 2010

Available online 25 September 2010

Received by Leonardo Marino-Ramirez

Keywords:

Animal model
MHC class II molecule
Molecular evolution
New world monkeys
Platyrrhini

ABSTRACT

The *Aotus nancymae* (owl monkey) is an important animal model in biomedical research, particularly for the preclinical evaluation of vaccine candidates against *Plasmodium falciparum* and *Plasmodium vivax*, which require a precisely typed major histocompatibility complex. The exon 2 from *A. nancymae* MHC-DPA1 gene was characterised in order to infer its allelic diversity and evolutionary history. Aona-DPA1 shows no polymorphism and is related to other primate DPA alleles (including Catarrhini and Platyrrhini), constituting an ancient trans-specific and strongly supported lineage with different variability and selective patterns when compared to other primate-MHC-DPA1 lineages. *A. nancymae* monkeys have thus a smaller MHC-DP polymorphism than MHC-DQ or MHC-DR.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Major histocompatibility complex (MHC) class II molecules display peptides on the surface of antigen-presenting cells (APC) for subsequent recognition by T cells, thereby performing a key defence role against pathogens. MHC class II molecules are heterodimers assembled from an α and a β glycopeptide chains encoded by the MHC class II A and B genes, respectively. Three main MHC class II loci, named HLA-DR, -DQ, and -DP, encode functional antigen-presenting molecules in primates. Genetic polymorphism and diversifying selection tied to functional and structural restrictions are common characteristics of these main loci. Such polymorphism is mainly restricted to the second exon of MHC class II A and B genes, constituting the molecule's peptide binding region (PBR) (Klein et al., 1993b).

MHC-DP is an ancient locus shared by divergent mammalian orders (Takahashi et al., 2000; Yuhki et al., 2003). However, its polymorphism and functionality vary. For example, MHC-DP acquires

a pseudo-genic nature in felines, as also occurs in murinae (mouse-like rodents), even though MHC-DP is the most polymorphic MHC class II locus in other rodents, such as the mole rat (*Spalax* genus) (Klein et al., 1993a; Yuhki et al., 2003; Kelley et al., 2005).

MHC-DP is the most centromeric locus within the primate MHC gene cluster region, being constituted by four genes: DPA1 and DPB1 genes and DPA2 and DPB2 pseudogenes. This arrangement (position and number) is apparently the same in all primates and was established before the split between Platyrrhini and Catarrhini ~43 million years ago (MY) (Klein et al., 1993a; Steiper and Young, 2006).

MHC-DPA1 variability in primates varies amongst nonexistent and low polymorphism, whilst for MHC-DPB1, it fluctuates from moderate to high polymorphism (Otting and Bontrop, 1995; Slierendregt et al., 1995; Bontrop et al., 1999; Doxiadis et al., 2001). HLA-DPA1 exhibits low polymorphism in humans, where 28 alleles have been reported to date, compared to the 138 alleles described for HLA-DPB1 (Robinson et al., 2003). In contrast, *Callithrix jacchus* (the common marmoset, a neo-tropical primate), has the MHC-DP region inactive, not expressing any MHC-DP molecule (Antunes et al., 1998). In spite of such low polymorphism, MHC-DPA1 can be important in modulating an immune response, since HLA-DPA1*0301 appears to be involved in the genetic susceptibility to *Schistosoma haematobium* and several chronic inflammatory diseases (May et al., 1998; Dai et al., 2010).

Previous studies have characterised *Aotus* MHC class II genes and molecules: MHC DQA-DQB (Diaz et al., 2000), MHC-DRB1 (Niño-Vasquez et al., 2000; Suarez et al., 2006), and MHC-DPB1 (Diaz et al., 2002). These neo-tropical primates have been shown to be susceptible to various human infectious diseases (Lujan et al., 1986; Polotsky

Abbreviations: MHC, Major histocompatibility complex; APC, antigen-presenting cells; PBR, peptide binding region; NWM, new world monkeys; OWM, old world monkeys; NJ, neighbour joining; ME, minimum-evolution; ML, maximum likelihood; LRSH, local rearrangements of tree topology around an edge; Pars, parsimony; GRMD, global rate minimum deformation method; MY, million years; SLAC, single likelihood ancestor counting; FEL, fixed effects likelihood; REL, random effects likelihood; Sub/S/MY, substitution per site per million years; TSP, trans-specific polymorphism.

* Corresponding author. Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá, Colombia. Tel.: +57 1 4815219; fax: +57 1 4815269.

E-mail address: mepatar@gmail.com (M.E. Patarroyo).

Grupos de Investigación en Colombia

1. Líder de Grupo: Genética Molecular Vegetal, Biología Computacional y Bioinformática. Código del grupo (COLCIENCIAS): COL0085459

Clasificación 2009: A

Clasificación 2010: A

2. Participante: MICOBAC-UN. Código del grupo (COLCIENCIAS): COL0078428

Clasificación 2010: A1

Genética Molecular Vegetal, Biología Computacional y Bioinformática

Datos básicos

Año y mes de formación	2003 - 1
Departamento - Ciudad	Cundinamarca - Mosquera
Líder	Leonardo Mariño-Ramírez
¿La información de este grupo se ha certificado?	Si el día 2008-12-03
Página web	
E-mail	lmario@corpoica.org.co
Clasificación	A
Área de conocimiento	Ciencias Biológicas -- Genética
Programa nacional de ciencia y tecnología	Biotecnología
Programa nacional de ciencia y tecnología (secundario)	Biotecnología

Instituciones

1.- Corporación Colombiana De Investigación Agropecuaria - Corpoica - (Avalado)

Líneas de investigación declaradas por el grupo

- 1.- Bioinformática
- 2.- Biología Computacional
- 3.- Genética de ganado criollo
- 4.- Genómica Funcional
- 5.- Genómica de Microorganismos
- 6.- Mejoramiento genético
- 7.- Propagación In Vitro

Sectores de aplicación

Integrantes del grupo

Nombre	Vinculación	Horas dedicación	Inicio - Fin Vinculación
1.- Leonardo Mariño-Ramírez	Investigador	20	2003/1 - Actual
2.- Luz Stella Barrero Meneses	Investigador	0	2007/1 - Actual
3.- Evelyn Gisela Arenas Ochoa	Investigador	0	2009/1 - Actual
4.- Oscar Camilo Bedoya Reina	Investigador	0	2007/3 - Actual
5.- Felix Enciso Rodriguez	Investigador	40	2008/1 - Actual
6.- Gina Garzón Martínez	Investigador	40	2009/1 - Actual
7.- Linda Yhiset Gómez Arias	Investigador	40	2008/1 - Actual
8.- Silvia Gómez Daza	Investigador	0	2007/1 - Actual
9.- Irving King Jordan	Investigador		2003/1 - Actual
10.- Irving King Jordan	Investigador	0	2007/1 - Actual
11.- David Landsman	Investigador	0	2007/3 - Actual
12.- Víctor Manuel Núñez Zarantes	Investigador	0	2007/1 - Actual
13.- Jaime Simbaqueba Gonzalez	Investigador	40	2008/1 - Actual
14.- John Spouge	Investigador	0	2007/3 - Actual

15.- Erika Patricia Sánchez Betancourt	Investigador	40	2008/1 - Actual
16.- ANA MILENA VALDERRAMA FONSECA	Investigador	0	2007/1 - Actual
17.- Ana Milena Valderrama Fonseca	Investigador	0	2007/1 - Actual
18.- Sandra Patricia Valbuena Aguilar	Estudiante	40	2008/1 - 2008/8

Producción

Artículos publicados en revistas científicas

- 1.- **Completo:** Producción de líneas homocigotas de arroz tolerantes a la toxicidad de aluminio mediante el cultivo de anteras.
Colombia, Arroz ISSN: 0, 1985 vol:34 fasc: 337 págs: 12 - 17
Autores: VICTOR MANUEL NUNEZ ZARANTES, 4NUNEZ VMJ NARVAEZ CP MARTINEZ AND W ROCA,
- 2.- **Completo:** Importance of Anther Culture in Rice Breeding
Colombia, Arroz ISSN: 0, 1987 vol:34 fasc: 337 págs: 12 - 17
Autores: VICTOR MANUEL NUNEZ ZARANTES, MARTINEZ CP VM NUNEZ AND W ROCA,
- 3.- **Completo:** A SCAR marker for the sex types determination in Colombian genotypes of Carica papaya
Holanda, Euphytica ISSN: 0014-2336, 2007 vol:153 fasc: págs: 215 - 220
Autores: VICTOR MANUEL NUNEZ ZARANTES, GIOVANNI CHAVES BEDOYA,
- 4.- **Completo:** Estudio preliminar para el establecimiento de un protocolo de criopreservación para palma de aceite (Elaeis guineensis Jacq)
Colombia, Agronomía Colombiana ISSN: 0120-9965, 2007 vol:25 fasc: 2 págs: 215 - 223
Autores: ALBA LUCIA VILLA, PABLO EDGAR JIMENEZ, RAUL IVAN VALBUENA, SILVIO BASTIDAS, VICTOR MANUEL NUNEZ ZARANTES,
- 5.- **Completo:** Evaluación de la toxicidad de proteínas de Bacillus thuringiensis Berliner sobre el picudo del algodón Anthonomus grandis Boheman
Colombia, Agronomía Colombiana ISSN: 0120-9965, 2006 vol:24 fasc: 2 págs: 296 - 301
Autores: SYLVIA GOMEZ, GUSTAVO DIAZ, VICTOR MANUEL NUNEZ ZARANTES,
- 6.- **Completo:** Recombinant Cry3Aa has insecticidal activity against the andean potato weevil, Premnotrypes vorax.
Estados Unidos, Biochemical And Biophysical Research Communications ISSN: 0006-291X, 2000 vol:279 fasc: págs: 653 - 656
Autores: SILVIA GOMEZ DAZA, CONSTANZA MATEUS, JAVIER HERNANDEZ, BARBARA ZIMMERMAN,
- 7.- **Completo:** Identificación de genes R1 y R2 que contienen resistencia a Phytophthora infestans en genotipos colombianos de papa
Colombia, Revista Colombiana De Biotecnología ISSN: 0123-3475, 2003 vol:5 fasc: 2 págs: 40 - 50
Autores: MARCELA DIAZ, DIEGO FAJARDO, JOSE DILMER MORENO, CELSA GARCIA, VICTOR MANUEL NUNEZ ZARANTES,
- 8.- **Completo:** Evaluación de la toxicidad de proteínas de Bacillus thuringiensis Berliner sobre el picudo del algodón Anthonomus grandis Boheman
Colombia, Agronomía Colombiana ISSN: 0120-9965, 2006 vol:24 fasc: 2 págs: 296 - 301
Autores: SILVIA GOMEZ DAZA, GUSTAVO DIAZ, VICTOR MANUEL NUNEZ,
- 9.- **Completo:** Genetic variability of Beauveria bassiana associated with the Coffee Berry Borer Hypothenemus hampei and other insects.
Colombia, Mycological Research ISSN: 0953-7562, 2002 vol:106 fasc: 11 págs: 1307 - 1314
Autores: ANA MILENA VALDERRAMA FONSECA, ALVARO GAITAN, GABRIEL SILDARRIAGA, PATRICIA VELEZ, ALEX BUSTILLO,
- 10.- **Completo:** Aplicación de la proteína verde fluorescente para el monitoreo de cepas degradadoras de fenol.
Colombia, Revista Colombiana De Biotecnología ISSN: 0123-3475, 2001 vol:III fasc: 2 págs: 78 - 84
Autores: ANA MILENA VALDERRAMA FONSECA, JULIA RAQUEL ACERO,
- 11.- **Completo:** Análisis de la variabilidad genética del hongo entomopatógeno Beauveria bassiana con marcadores RAPD
Colombia, Revista de La Sociedad Colombiana de Entomología ISSN: 0, 2000 vol:26 fasc: 1-2 págs: 25 - 29
Autores: ANA MILENA VALDERRAMA FONSECA, MARCO AURELIO CRISTANCHO, BERNARDO CHAVES,
- 12.- **Completo:** Caracterización de aislamientos de Beauveria bassiana para el control de la broca del café
Colombia, Manejo Integrado De Plagas ISSN: 1016-0469, 2001 vol:62 fasc: págs: 38 - 53
Autores: ANA MILENA VALDERRAMA FONSECA, NANCY ESTRADA, MARIA TERESA GONZALES, PATRICIA VELEZ, ALEX BUSTILLO,
- 13.- **Completo:** Regulatory Change in YABBY-like Transcription Factor Led to Evolution of Extreme Fruit Size during Tomato Domestication

Inglaterra, Nature Genetics ISSN: 1061-4036, 2008 vol:40 fasc: 6 págs: 800 - 804

Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY, BIN CONG,

14.- **Completo:** Evaluating the genetic basis of multiple locule fruit in a broad cross section of tomato cultivars
Alemania, Theoretical And Applied Genetics ISSN: 0040-5752, 2004 vol:109 fasc: 3 págs: 669 - 679

Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY,

15.- **Completo:** Allelism test among high locule number tomato mutants and genetic mapping of the loci involved
Estados Unidos, Journal Of Experimental Botany ISSN: 1460-2431, 2001 vol:51 fasc: págs: 11 - 13

Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY,

16.- **Revista (magazín):** Towards the development of a basic genomics platform for exotic fruit solanaceae
Estados Unidos, SOL newsletter ISSN: 0, 2005 vol: fasc: págs: -

Autores: LUZ STELLA BARRERO MENESES,

17.- **Completo:** Developmental characterization of the fasciated locus and mapping of Arabidopsis candidate genes involved in the control of floral meristem size and carpel number in tomato

Canadá, Genome ISSN: 1480-3321, 2006 vol:49 fasc: págs: 991 - 1006

Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY, FEINAN WU, BIN CONG,

18.- **Completo:** Construcción de bancos de genes y caracterización de germoplasma

Colombia, Agricultura De Las Americas ISSN: 0120-6052, 2005 vol: fasc: 347 págs: 44 - 47

Autores: LUZ STELLA BARRERO MENESES,

19.- **Completo:** Transposable element derived DNaseI-hypersensitive sites in the human genome.

Inglaterra, Biology International ISSN: 1745-6150, 2006 vol:1 fasc: 20 págs: 1 -

Autores: LEONARDO MARINO-RAMIREZ, IRVING KING JORDAN,

20.- **Completo:** Expression patterns of protein kinases correlate with gene architecture and evolutionary rates

Estados Unidos, Plos One ISSN: 1932-6203, 2008 vol:3 fasc: 10 págs: e3599 -

Autores: ALEKSEY OGURTSOV, GIBBES JOHNSON, SVETLANA SHABALINA, NIKOLAY SPIRIDONOV, LEONARDO MARINO-RAMIREZ, DAVID LANDSMAN,

21.- **Completo:** Identification and mapping of self-assembling protein domains encoded by the Escherichia coli K-12 genome using lambda repressor fusions.

Estados Unidos, Journal Of Bacteriology ISSN: 0021-9193, 2004 vol:186 fasc: págs: 1311 - 1319

Autores: LEONARDO MARINO-RAMIREZ, JAMES HU, JONATHAN MINOR, NICOLA READING,

22.- **Completo:** Origin and evolution of human microRNAs from transposable elements.

Estados Unidos, Genetics ISSN: 0016-6731, 2007 vol:176 fasc: págs: 1323 - 1337

Autores: JITTIMA PIRIYAPONGSA, LEONARDO MARINO RAMIREZ, IRVING KING JORDAN,

23.- **Completo:** Isolation and mapping of self-assembling protein domains encoded by the Saccharomyces cerevisiae genome.

Estados Unidos, Yeast ISSN: 0749-503X, 2002 vol:19 fasc: págs: 641 - 650

Autores: LEONARDO MARINO-RAMIREZ, JAMES HU,

24.- **Completo:** Uso de la reacción en cadena de la polimerasa para la caracterización de aislamientos nativos de Bacillus thuringiensis.

Colombia, Revista Corpoica - Ciencia Y Tecnologia Agropecuarias ISSN: 0122-8706, 1996 vol:2 fasc: págs: 2 - 9

Autores: LEONARDO MARINO-RAMIREZ,

25.- **Corto (Resumen):** Caracterización Molecular de Genes cry de Bacillus thuringiensis utilizando PCR Extra-Rápida.

Colombia, Revista Corpoica - Ciencia Y Tecnologia Agropecuarias ISSN: 0122-8706, 1996 vol:2 fasc: págs: 47 - 47

Autores: LEONARDO MARINO-RAMIREZ,

26.- **Completo:** Caracterización del gen de la proteína de la cápside de dos aislamientos del virus del mosaico del pepino (CMV), obtenidos de plátano y banano (Musa spp.).

Colombia, Revista Corpoica - Ciencia Y Tecnologia Agropecuarias ISSN: 0122-8706, 1996 vol:1 fasc: págs: 1 - 5

Autores: LEONARDO MARINO-RAMIREZ,

27.- **Completo:** Global regulation by the yeast Spt10 protein is mediated through chromatin structure and the histone upstream activating sequence elements.

Estados Unidos, Molecular And Cellular Biology ISSN: 0270-7306, 2005 vol:25 fasc: 20 págs: 9127 - 9137

Autores: PETER ERIKSSON, GEETU MENDIRATTA, NEIL MCLAUGHLIN, TYRA WOLFSBERG, LEONARDO MARINO-RAMIREZ, TIFFANY POMPA, MOHENDRA JAINERIN, DAVID LANDSMAN, CHANG HUI SHEN, DAVID CLARK,

28.- **Completo:** Transposable elements donate lineage-specific regulatory sequences to host genomes.

Suiza, Cytogenetic And Genome Research - Online ISSN: 1424-8581, 2005 vol:110 fasc: 1 págs: 333 - 341

Autores: LEONARDO MARINO-RAMIREZ, DAVID LANDSMAN, IRVING KING JORDAN, KEVIN LEWIS,

29.- **Completo:** TLX1/HOX11-induced hematopoietic differentiation blockade.

Inglaterra, Oncogene ISSN: 0950-9232, 2007 vol:26 fasc: págs: 4115 - 4123

Autores: DAVID LANDSMAN, I RIZ, SERGEY AKIMOV, SHANNON EAKER, KK BAXTER, H LEE, LEONARDO MARINO-RAMIREZ, TERESA HAWLEY, ROBERT HAWLEY,

30.- **Completo:** Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. Inglaterra, Bioinformatics ISSN: 1367-4803, 2005 vol:Suppl1 fasc: págs: i440 - i448
Autores: LEONARDO MARINO-RAMIREZ, DAVID LANDSMAN, KANNAN THARAKARAMAN, SERGEY SHEETLIN, JOHN SPOUGE,

31.- **Corto (Resumen):** Clonación del gen de la cápside proteica de una cepa colombiana del virus del mosaico del pepino (CMV) para su expresión en plantas por transformación mediante Agrobacterium Colombia, Revista Corpoica - Ciencia Y Tecnología Agropecuarias ISSN: 0122-8706, 1997 vol:2 fasc: págs: 58 - 59
Autores: LEONARDO MARINO-RAMIREZ,

32.- **Completo:** Evolutionary significance of gene expression divergence. Holanda, Gene ISSN: 0378-1119, 2005 vol:17 fasc: 345 págs: 119 - 126
Autores: LEONARDO MARINO-RAMIREZ, IRVING KING JORDAN, EUGENE KOONIN,

33.- **Completo:** Expression patterns of protein kinases correlate with gene architecture and evolutionary rates Estados Unidos, Plos One ISSN: 1932-6203, 2008 vol:3 fasc: 10 págs: e3599 -
Autores: LEONARDO MARINO RAMIREZ, GIBBES JOHNSON, SVETLANA SHABALINA, NIKOLAY SPIRIDONOV, DAVID LANDSMAN, ALEKSEY OGURTSOV,

34.- **Completo:** Database resources of the National Center for Biotechnology Information Inglaterra, Nucleic Acids Research ISSN: 0305-1048, 2008 vol:37 fasc: 1 págs: D5 - D15
Autores: DAVID LANDSMAN, ERIC SAYERS,

35.- **Completo:** TLX1/HOX11-induced hematopoietic differentiation blockade. Inglaterra, Oncogene ISSN: 0950-9232, 2007 vol:26 fasc: págs: 4115 - 4123
Autores: DAVID LANDSMAN, TERESA HAWLEY, ROBERT HAWLEY, LEONARDO MARINO RAMIREZ, I RIZ, SERGEY AKIMOV, SHANNON EAKER, KK BAXTER, H LEE,

36.- **Completo:** Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. Inglaterra, Bmc Genomics ISSN: 1471-2164, 2008 vol:17 fasc: págs: 226 -
Autores: NALINI POLAVARAPU, LEONARDO MARINO RAMIREZ, DAVID LANDSMAN, JOHN MCDONALD, IRVING KING JORDAN,

37.- **Completo:** Transposable elements donate lineage-specific regulatory sequences to host genomes. Suiza, Cytogenetic And Genome Research - Online ISSN: 1424-8581, 2005 vol:110 fasc: págs: 333 - 341
Autores: LEONARDO MARINO RAMIREZ, KEVIN LEWIS, DAVID LANDSMAN, IRVING KING JORDAN,

38.- **Completo:** Co-evolutionary Rates of Functionally Related Yeast Genes. Nueva Zelanda, Evolutionary Bioinformatics ISSN: 1176-9343, 2006 vol:2 fasc: págs: 295 - 300
Autores: LEONARDO MARINO RAMIREZ, OLIVIER BODENREIDER, NATALIE KANTZ, IRVING KING JORDAN,

39.- **Completo:** Transposable element derived DNaseI-hypersensitive sites in the human genome. Inglaterra, Biology International ISSN: 1745-6150, 2006 vol:20 fasc: págs: 1 - 20
Autores: IRVING KING JORDAN, LEONARDO MARINO RAMIREZ,

40.- **Completo:** Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. Inglaterra, Bioinformatics ISSN: 1367-4803, 2005 vol:Suppl1 fasc: págs: i440 - i448
Autores: KANNAN THARAKARAMAN, LEONARDO MARINO RAMIREZ, SERGEY SHEETLIN, DAVID LANDSMAN, JOHN SPOUGE,

41.- **Completo:** Statistical analysis of over-represented words in human promoter sequences. Inglaterra, Nucleic Acids Research ISSN: 0305-1048, 2004 vol:12 fasc: 32 págs: 949 - 958
Autores: JOHN SPOUGE, LEONARDO MARINO RAMIREZ, DAVID LANDSMAN, GAVIN KANGA,

42.- **Completo:** The Histone Database: a comprehensive resource for histones and histone fold-containing proteins. Estados Unidos, Proteins-Structure Function And Bioinformatics ISSN: 0887-3585, 2006 vol:1 fasc: 62 págs: 838 - 842
Autores: LEONARDO MARINO RAMIREZ, BENJAMIN HSU, ANDREAS BAXEVANIS, DAVID LANDSMAN,

43.- **Completo:** Multiple independent evolutionary solutions to core histone gene regulation. Inglaterra, Genome Biology ISSN: 1465-6906, 2006 vol:7 fasc: págs: R122 -
Autores: LEONARDO MARINO RAMIREZ, IRVING KING JORDAN, DAVID LANDSMAN,

44.- **Revisión (Survey):** Histone structure and nucleosome stability. Inglaterra, Expert Review Of Proteomics ISSN: 1478-9450, 2005 vol:2 fasc: 5 págs: 719 - 729
Autores: LEONARDO MARINO RAMIREZ, MARICEL KANN, BENJAMIN SHOEMAKER, DAVID LANDSMAN,

45.- **Completo:** Global regulation by the yeast Spt10 protein is mediated through chromatin structure and the histone upstream activating sequence elements. Estados Unidos, Molecular And Cellular Biology ISSN: 0270-7306, 2005 vol:25 fasc: págs: 9127 - 9137
Autores: PETER ERIKSSON, GEETU MENDIRATTA, NEIL MCLAUGHLIN, TYRA WOLFSBERG, LEONARDO MARINO RAMIREZ, TIFFANY POMPA, MOHENDRA JAINERIN, DAVID LANDSMAN, CHANG HUI SHEN, DAVID CLARK,

- 46.- **Completo:** Global similarity and local divergence in human and mouse gene co-expression networks. Inglaterra, Bmc Evolutionary Biology ISSN: 1471-2148, 2006 vol:12 fasc: págs: 6 - 70
Autores: PANAYIOTIS TSAPARAS, LEONARDO MARINO RAMIREZ, OLIVIER BODENREIDER, EUGENE KOONIN, IRVING KING JORDAN,
- 47.- **Completo:** Conservation and coevolution in the scale-free human gene coexpression network. Estados Unidos, Molecular Biology And Evolution ISSN: 0737-4038, 2004 vol:21 fasc: págs: 2058 - 2070
Autores: IRVING KING JORDAN, YURI WOLF, LEONARDO MARINO RAMIREZ, EUGENE KOONIN,
- 48.- **Completo:** Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. Inglaterra, Bmc Bioinformatics ISSN: 1471-2105, 2008 vol:4 fasc: págs: 262 -
Autores: NAK KYEONG KIM, KANNAN THARAKARAMAN, LEONARDO MARINO RAMIREZ, JOHN SPOUGE,
- 49.- **Completo:** Scanning sequences after Gibbs sampling to find multiple occurrences of functional elements. Inglaterra, Bmc Bioinformatics ISSN: 1471-2105, 2006 vol:8 fasc: págs: 408 -
Autores: KANNAN THARAKARAMAN, LEONARDO MARINO RAMIREZ, SERGEY SHEETLIN, DAVID LANDSMAN, JOHN SPOUGE,
- 50.- **Completo:** The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. Inglaterra, Nucleic Acids Research ISSN: 0305-1048, 2008 vol:36 fasc: págs: 2777 - 2786
Autores: KANNAN THARAKARAMAN, OLIVIER BODENREIDER, DAVID LANDSMAN, JOHN SPOUGE, LEONARDO MARINO RAMIREZ,

Trabajos en eventos (Capítulos de memoria)

- 1.- **Completo** : Florula de la reserva forestal protectora el Malmo Colombia, Evento: Encuentro nacional de semilleros de investigación Ponencia: año:2004, Resúmenes Encuentro nacional de semilleros de investigación ISBN: 0 vol: págs: ,
Autores: FELIX ENCISO RODRIGUEZ,
- 2.- **Completo** : Caracterización Molecular del lulo (Solanum quitoense) y tomate de árbol (Solanum betaceum) del banco de germoplasma de Corpoica mediante el empleo de marcadores COS II Colombia, Evento: X Congreso de la asociación colombiana de fitomejoramiento y producción de cultivos Ponencia: año:2007, Asociación colombiana de fitomejoramiento y producción de cultivos ISBN: 0 vol: págs: ,
Autores: FELIX ENCISO RODRIGUEZ,
- 3.- **Resumen** : Ab Initio prediction of Transcription Factor Binding Sites in intergenic spacers of histone core genes. Estados Unidos, Evento: 20th NIH Research Festival Ponencia: año:2007, ISBN: vol: págs: ,
Autores: OSCAR CAMILO BEDOYA REINA, LEONARDO MARINO RAMIREZ,
- 4.- **Resumen** : Aplicación de un método para la detección de microorganismos nitrificantes y su evaluación frente a diferentes parámetros ambientales México, Evento: IV Congreso de Biotecnología y Bioingeniería Ponencia: año:1999, Resúmenes del IV Congreso de Biotecnología y Bioingeniería ISBN: 0 vol: págs: ,
Autores: ANA MILENA VALDERRAMA FONSECA, JULIA RAQUEL ACERO,
- 5.- **Resumen** : Transformación genética de papa colombiana con genes cry Colombia, Evento: IX Congreso de la Corporación para Investigaciones Biológicas Ponencia: año:2003, Resúmenes del IX Congreso de la Corporación para Investigaciones Biológicas ISBN: 0 vol: págs: ,
Autores: ANA MILENA VALDERRAMA FONSECA, RAFAEL ARANGO ISAZA, ESPERANZA RODRIGUEZ,
- 6.- **Resumen** : Construcciones genéticas cry1Ab y cry1Ac de Bacillus thuringiensis para el desarrollo de líneas de papa con posible resistencia a Tectia solanivora Colombia, Evento: VIII Congreso de la Corporación para Investigaciones Biológicas Ponencia: año:2002, Resúmenes del VIII Congreso de la Corporación para Investigaciones Biológicas ISBN: 0 vol: págs: ,
Autores: ANA MILENA VALDERRAMA FONSECA, RAFAEL ARANGO ISAZA, ESPERANZA RODRIGUEZ,
- 7.- **Resumen** : XXVII Congreso de la Sociedad Colombiana de Entomología Colombia, Evento: XXVII Congreso de la Sociedad Colombiana de Entomología Ponencia: año:2000, Resúmenes del XXVII Congreso de la Sociedad Colombiana de Entomología ISBN: 0 vol: págs: ,
Autores: ANA MILENA VALDERRAMA FONSECA, ALVARO GAITAN, GABRIEL SILDARRIAGA, PATRICIA VELEZ, ALEX BUSTILLO,
- 8.- **Resumen** : Análisis de la variabilidad genética del hongo entomopatógeno Beauveria bassiana con marcadores RAPD Colombia, Evento: XXV Congreso de la Sociedad Colombiana de Entomología Ponencia: año:1998, Resúmenes del XXV Congreso de la Sociedad Colombiana de Entomología ISBN: 0 vol: págs: ,
Autores: ANA MILENA VALDERRAMA FONSECA, PATRICIA VELEZ, ALEX BUSTILLO,
- 9.- **Completo** : Application of genomics tools - case study tomato - Solanaceae Estados Unidos, Evento: American Society of Plant Biologists Ponencia: año:2008, Proceedings of the American Society of

Plant Biologists (ASPB) 2008 ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES,

10.- **Completo** : Statistical Analysis of DNA Sequences in Human Promoter Regions
Estados Unidos, Evento: 2003 NIH Research Festival Ponencia: año:2003, NIH Press ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,

11.- **Resumen** : Evaluating the genetic and molecular bases of multiple locule fruit in tomato
Estados Unidos, Evento: Plant & Animal Genomes XII Conference Ponencia: año:2004, Proceedings of the Plant & Animal Genomes XII Conference ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY,

12.- **Resumen** : Towards the development of a basic genomics platform for exotic fruit solanaceae
Italia, Evento: Second Solanaceae Genome Workshop 2005 Ponencia: año:2005, Proceedings of the Second Solanaceae Genome workshop ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES,

13.- **Resumen** : Evaluación de la base genética y molecular del numero de lóculos en tomate
Colombia, Evento: 9 Congreso de la Sociedad Colombiana de Fitomejoramiento y Produccion de Cultivos Ponencia: año:2005, Resúmenes del 9 Congreso de la Sociedad Colombiana de Fitomejoramiento y Produccion de Cultivos ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY,

14.- **Resumen** : Use of COS markers for the Andean fruited species lulo and tree tomato
Estados Unidos, Evento: PAA/Solanaceae 2006. Genomics meets biodiversity Ponencia: año:2006, Proceedings of the PAA/Solanaceae 2006. Genomics meets biodiversity ISBN: 0 vol: págs: ,
Autores: ALEXANDRA CRISTINA OLARTE, MARIO LOBO, STEVEN D TANKSLEY, LUZ STELLA BARRERO MENESES,

15.- **Completo** : Análisis comparativo de los perfiles electroforéticos de proteínas totales de las especies aceitera *Elaeis guineensis* y *E. oleifera*
Colombia, Evento: V congreso de la sociedad colombiana de fitomejoramiento y producción de cultivos Ponencia: año:1997, Memorias del V congreso de la sociedad colombiana de fitomejoramiento y produccion de cultivos ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES, JAVIER NARVAEZ, VIVIAN HERNANDEZ,

16.- **Resumen** : Marcadores bioquímicos en un programa de introgresion rapida de genes entre las especies aceiteras *Elaeis guineensis* y *E. oleifera*
Colombia, Evento: II Latin American Meeting of Plant Biotechnology / REDBIO Ponencia: año:1995, Proceedings of the II Latin American Meeting of Plant Biotechnology / REDBIO ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES, JAVIER NARVAEZ, SILVIO BASTIDAS, RAFAEL REYES,

17.- **Resumen** : Marcadores bioquímicos en un programa de introgresion rapida de genes entre las especies aceiteras *Elaeis guineensis* y *E. oleifera*
Colombia, Evento: X Convención Científica Colombiana Ponencia: año:1995, Memorias de la X Convención Científica Colombiana ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES, JAVIER NARVAEZ, SILVIO BASTIDAS, RAFAEL REYES,

18.- **Completo** : Marcadores bioquímicos en un programa de introgresión rápida de genes entre las especies aceiteras *Elaeis guineensis* y *E. oleifera*
Colombia, Evento: IV congreso de la sociedad colombiana de fitomejoramiento y produccion de cultivos Ponencia: año:1995, Memorias del IV congreso de la sociedad colombiana de fitomejoramiento y producción de cultivos ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES, JAVIER NARVAEZ, SILVIO BASTIDAS, RAFAEL REYES,

19.- **Completo** : Identificación y aislamiento de genes de defensa a Sigatoka negra de germoplasma de banano
Colombia, Evento: V congreso de la sociedad colombiana de fitomejoramiento y producción de cultivos Ponencia: año:1997, Memorias del V congreso de la sociedad colombiana de fitomejoramiento y produccion de cultivos ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES, JAVIER NARVAEZ, LEONARDO MARINO, NESON TORO, JUAN CARLOS POLANCO,

20.- **Resumen** : Genetic characterization of major high locule number loci in tomato
Estados Unidos, Evento: Plant, Animal & Microbe Genomes X Conference Ponencia: año:2002, Proceedings of the Plant, Animal & Microbe Genomes X Conference ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY,

21.- **Resumen** : Mapping and Isolation of the fasciated locus: A major fruit locule number locus in tomato
Estados Unidos, Evento: Plant & Animal Genomes XIV Conference Ponencia: año:2006, Proceedings of the Plant & Animal Genomes XIV Conference ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY, BIN CONG,

22.- **Completo** : Caracterización molecular de lulo y tomate de árbol del banco de germoplasma de Corpoica mediante el empleo de marcadores COSII

- Colombia, Evento: X Congreso de la Asociación Colombiana de Fitomejoramiento y producción de Cultivos Ponencia: año:2007, Memorias del X Congreso de la Asociación Colombiana de Fitomejoramiento y producción de Cultivos ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES, FELIX EUGENIO ENCISO,
- 23.- **Resumen** : Avances en el uso de marcadores moleculares COSII para identificación de híbridos y mapeo genético comparativo en lulo
Colombia, Evento: X Congreso de la Asociación Colombiana de Fitomejoramiento y producción de Cultivos Ponencia: año:2007, Memorias del X Congreso de la Asociación Colombiana de Fitomejoramiento y producción de Cultivos ISBN: 0 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES, ZULMA CARDENAS,
- 24.- **Completo** : Genómica estructural y comparativa: Caso solanáceas-tomate
Colombia, Evento: IV Congreso Internacional y VII Congreso Colombiano de Genética Ponencia: año:2006, Salud. Revista de la Facultad de Salud Universidad Industrial de Santander ISBN: 0121080 vol: págs: ,
Autores: LUZ STELLA BARRERO MENESES,
- 25.- **Completo** : The precise positioning of genomic landmarks can aid in the identification of regulatory elements
Estados Unidos, Evento: International Society for Computational Biology (ISMB) meeting 2005 Ponencia: año:2005, (ISMB) meeting 2005 ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,
- 26.- **Completo** : Identification of positionally clustered sequence elements in human proximal promoter regions
Estados Unidos, Evento: 2004 Genomics Workshop on Identification of Functional Elements in Mammalian Genomes Ponencia: año:2004, Cold Spring Harbor Laboratory Press ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,
- 27.- **Completo** : Sequence evolution and the human gene expression network
Estados Unidos, Evento: Genomes and Evolution 2004 Ponencia: año:2004, The Pennsylvania State University Press ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,
- 28.- **Completo** : Mycobacterium tuberculosis protein oligomerization domains identified with repressor fusions
Estados Unidos, Evento: Third TB Structural Genomics Consortium Retreat Ponencia: año:2002, Third TB Structural Genomics Consortium Proceedings ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,
- 29.- **Completo** : Genome-wide mapping of homotypic interactions encoded in microbial genomes
Estados Unidos, Evento: Molecular Genetics of Bacteria & Phages Meeting Ponencia: año:2003, University of Wisconsin Press ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,
- 30.- **Completo** : Identification of candidate regulatory sequence elements in Homo sapiens
Estados Unidos, Evento: LXVIII Cold Spring Harbor Symposium on Quantitative Biology Ponencia: año:2003, Cold Spring Harbor Laboratory Press ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,
- 31.- **Completo** : Characterization of protein homotypic interactions encoded by four microbial genomes
Estados Unidos, Evento: Lost Pines Molecular Biology Conference Ponencia: año:2002, Lost Pines Molecular Biology Conference Proceedings 2002 ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,
- 32.- **Completo** : Genome-wide mapping of protein oligomerization domains in microorganisms
Estados Unidos, Evento: Molecular Genetics of Bacteria & Phages 2002 Ponencia: año:2002, Cold Spring Harbor Laboratory Press ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,
- 33.- **Completo** : An integrated genomics database for beef cattle
Estados Unidos, Evento: Plant, Animal & Microbe Genomes X Conference Ponencia: año:2002, Plant, Animal & Microbe Genomes X Conference Proceedings ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,
- 34.- **Completo** : Genetic approaches to study protein interactions
Estados Unidos, Evento: The Association of Biomolecular Resource Facilities meeting Ponencia: año:2002, ABRF 2002 Proceedings ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,
- 35.- **Completo** : Genome-wide mapping of protein oligomerization domains in Saccharomyces cerevisiae
Estados Unidos, Evento: Beyond Genome 2001 Ponencia: año:2001, Beyond Genome 2001 Proceedings ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,
- 36.- **Completo** : Lambda repressor fusions as a tool to study protein oligomerization in Saccharomyces cerevisiae
Estados Unidos, Evento: Yeast Genetics and Molecular Biology Meeting 2000 Ponencia: año:2000, Yeast Genetics and

Molecular Biology Meeting 2000 Proceedings ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,

37.- **Completo** : Differential gene expression during the break in potato tuber dormancy
Canadá, Evento: Plant Biology Meeting ¿97 Ponencia: año:1997, Plant Biology Meeting ¿97 Proceedings ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,

38.- **Resumen** : Aislamiento y caracterización de cepas de Bacillus thuringiensis en Colombia
Colombia, Evento: XXII Congreso de la Sociedad Colombiana de Entomología SOCOLEN Ponencia: año:1995, Anales de Socolen 1995 ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,

39.- **Completo** : Molecular Characterization of Native Strains of Bacillus thuringiensis
Argentina, Evento: REDBIO ¿95: Second Latin-American Plant Biotechnology Meeting Ponencia: año:1995, REDBIO ¿95 Proceedings ISBN: 0 vol: págs: ,
Autores: LEONARDO MARINO-RAMIREZ,

Libros publicados

Capítulos de libro publicados

1.- **Capítulo de libro publicado** : CULTIVO DE TEJIDOS EN LA AGRICULTURA
Colombia, 1991, Cultivo De Tejidos En La Agricultura: Fundamentos, ISBN: 958-9183-15-8, Vol. , págs:79 - 77, Ed. Ciat
Autores: VICTOR MANUEL NUNEZ ZARANTES, L SZABADOS V M NUNEZ L M TELLO G MAFLA J ROA W M ROCA,

2.- **Capítulo de libro publicado** : CULTIVO DE ANTERAS EN EL MEJORAMIENTO DE PLANTAS
Colombia, 1991, Cultivo De Tejidos En La Agricultura: Fundamentos, ISBN: 958-9183-15-8, Vol. , págs:271 - 312, Ed. Ciat
Autores: VICTOR MANUEL NUNEZ ZARANTES, WVICTOR MANUEL NUNEZ ZARANTESM ROCA V M NUNEZ K MORNAN,

3.- **Capítulo de libro publicado** : Oat Haploid from Anther Culture and from Wide Hybridizations
Estados Unidos, 1995, In vitro Haploid Production in Higher Plants, ISBN: 0, Vol. , págs: - , Ed. Kluwer Academic Publishers
Autores: VICTOR MANUEL NUNEZ ZARANTES, RINES HE O RIERA V M NUNEZ D W DAVIS R L PHILLI,

4.- **Capítulo de libro publicado** : Monocotiledoneas
Colombia, 2006, Florula de la Reserva Forestal Protectora " El Malmo", ISBN: 0, Vol. Vol.1, págs:1 - 88, Ed. Publinprenta UPTC-Tunja
Autores: FELIX ENCISO RODRIGUEZ, JUAN CARLOS ZABALA, DANIEL HUMBERTO GALINDO, MARIA MARGARITA SUAREZ,

5.- **Capítulo de libro publicado** : Genomics of Tropical Solanaceous species: Established and emerging crops
Estados Unidos, 2008, Genomics of tropical Crop Plants: Plant Genetics and Genomics Crops and Models, ISBN: 0, Vol. 1, págs:453 - 467, Ed. Springer Verlag
Autores: RC PRATT, D FRANCIS, LUZ STELLA BARRERO MENESES,

6.- **Capítulo de libro publicado** : Using lambda repressor fusions to isolate and characterize self-assembling domains
Estados Unidos, 2002, Protein-Protein Interactions: A Laboratory Manual, ISBN: 0879696281, Vol. 1, págs:375 - 394, Ed. Cold Spring Harbor Laboratory Press
Autores: LEONARDO MARINO-RAMIREZ,

7.- **Capítulo de libro publicado** : Screening peptide/protein libraries fused to the lambda repressor DNA-binding Domain in E. coli cells
Estados Unidos, 2003, Methods in Molecular Biology: E. coli Gene Expression Protocols, ISBN: 1588290085, Vol. 205, págs:235 - 250, Ed. Humana Press, Inc.
Autores: LEONARDO MARINO-RAMIREZ,

8.- **Capítulo de libro publicado** : Evolutionary genomics of gene expression
Estados Unidos, 2007, Structural Approaches to Sequence Evolution, ISBN: 9783540353058, Vol. 1, págs:223 - 240, Ed. Springer
Autores: LEONARDO MARINO-RAMIREZ,

Textos en publicaciones no científicas

1.- **Revista (magazín)** : Towards the development of a basic genomics platform for exotic fruit solanaceae
Estados Unidos, SOL newsletter ISSN: 0, 2005 vol: fasc: págs: -
Autores: LUZ STELLA BARRERO MENESES,

Partituras musicales

Prefacio, epílogo

Otra producción bibliográfica

1.- **Documento de trabajo (working paper)** : Genetic and developmental characterization of locule number loci in tomato with emphasis on the fasciated locus.

Estados Unidos, 2004, , , vol. , págs: 1, - , Ed.

Autores: LUZ STELLA BARRERO MENESES,

2.- **Documento de trabajo (working paper)** : Aislamiento y caracterización parcial de un elicitor derivado de tubos germinales de la roya del café (Hemileia vastatrix Berk & Br.)

Colombia, 1994, , , vol. , págs: 1, - , Ed.

Autores: LUZ STELLA BARRERO MENESES,

Softwares

1.- **Computacional** : A-GLAM

Estados Unidos, 2005, , Irrestricita, UNIX/LINUX, , , Identificación de elementos de regulación en secuencias de ADN.

Autores: LEONARDO MARINO-RAMIREZ,

2.- **Computacional** : Doodle

Estados Unidos, 2002, , Irrestricita, UNIX/Linux, , , Plataforma para divulgar datos genómicos obtenidos en varios microorganismos.

Autores: LEONARDO MARINO-RAMIREZ,

Productos tecnológicos

1.- **Proyecto** : Caracterización de eventos del desarrollo floral afectados por el locus fasciated en tomate

Estados Unidos, 2006, , Irrestricita, Caracterizar eventos del desarrollo floral y vegetativo afectados por el locus fasciated en tomate.

Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY,

2.- **Proyecto** : Mil cuatrocientas secuencias parciales de genes COSII (Set de Ortólogos Conservados) en lulo (Solanum guineense), tomate de árbol (S. betaceum) y taxa relacionados (S. hirtum, S. uniloba)

Estados Unidos, 2006, , Irrestricita, Desarrollar bancos de genes en lulo y tomate de árbol.

Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY, MARIO LOBO, ALEXANDRA CRISTINA OLARTE,

3.- **Proyecto** : Secuencias de genes COSII en lulo (S. hirtum, S. guineense) y tomate de árbol (S. uniloba, S. betaceum)

Estados Unidos, 2006, , Irrestricita, Desarrollar bancos de genes en lulo y tomate de árbol.

Autores: ALEXANDRA CRISTINA OLARTE, MARIO LOBO, STEVEN D TANKSLEY, LUZ STELLA BARRERO MENESES,

4.- **Proyecto** : Cuatro genes candidatos para la característica número de lóculos del fruto del tomate identificados

Estados Unidos, 2006, , Irrestricita, Identificar genes candidatos para la característica número de lóculos en tomate.

Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY, FEINAN WU, BIN CONG,

5.- **Proyecto** : Seis características asociadas con el locus fasciated del tomate (tamaño del meristemo floral, número de órganos florales, tamaño y forma del ovario, tamaño y forma del fruto) en 5 estados del desarrollo floral

Estados Unidos, 2006, , Irrestricita, Caracterizar eventos del desarrollo floral y vegetativo afectados por el locus fasciated en tomate.

Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY,

6.- **Proyecto** : Catorce accesiones de tomate del Tomato Genetics Resources Center (TGRC, USA) caracterizadas para base genética (relaciones de alelismo) del número de lóculos del fruto

Estados Unidos, 2004, , Irrestricita, Identificar y caracterizar los loci del genoma del tomate que controlan el número de lóculos del fruto.

Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY,

7.- **Proyecto** : Once marcadores COSII polimórficos evaluados en 62 accesiones de lulo y tomate de árbol

Colombia, 2007, , Restricita, Caracterización molecular de colecciones de germoplasma nacionales.

Autores: LUZ STELLA BARRERO MENESES, FELIX EUGENIO ENCISO,

8.- **Proyecto** : Gen de tomate similar a factor de transcripción YABBY (fasciated), Números de accesión en GeneBank: EU577673 (ADN genómico), EU557674 (cDNA), EU557676 (alelo 1), EU557677 (promotor)

Estados Unidos, 2008, , Irrestricita, Controla número de lóculos y tamaño del fruto del tomate.

Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY, BIN CONG,

9.- **Proyecto** : Gen de tomate similar a proteína quinasa TOUSLED, Número de accesión en GeneBank: EU557675

Estados Unidos, 2008, , Irrestricita, Gen candidato que controla tamaño del fruto del tomate.

Autores: LUZ STELLA BARRERO MENESES, STEVEN D TANKSLEY, BIN CONG,

Procesos o técnicas

1.- **Otra** : RNA fingerprinting (DD-RTPCR) para la identificación de genes que se expresan diferencialmente en banano Colombia, 1996, , Restricta, Identificación de genes de banano.

Autores: LUZ STELLA BARRERO MENESES,

2.- **Otra** : Evaluación de diferentes Primers para la caracterización de M. fijiensis por RAPDs

Colombia, 1997, , Restricta, Evaluación de Primers.

Autores: LUZ STELLA BARRERO MENESES,

Trabajos técnicos

Normas

Cartas, mapas o similares

Cursos de corta duración dictados

Desarrollo de material didáctico o de instrucción

Editoración o revisión

Mantenimiento de obras artísticas

Organización de eventos

Programas en radio o TV

Informes de investigación

Presentación de trabajo

Otra producción técnica

Producción artística/cultural

Trabajos dirigidos/Tutorías concluidas

1.- **Trabajo de conclusión de curso de pregrado** : Caracterización bioquímica y molecular de cepas de *Bacillus thuringiensis* y evaluación biológica sobre larvas de *Anthonomus grandis* (picudo del algodón) Colombia, 2004, , Orientados: Gustavo Enrique Díaz Melendez, Ingeniería Agronómica, Universidad Nacional De Colombia. Autores: SILVIA GOMEZ DAZA,

2.- **Trabajo de conclusión de curso de pregrado** : Caracterización molecular y análisis de variabilidad genética en aislamientos de *Mycosphaerella fijiensis* provenientes de las regiones bananeras del Uraba y el eje cafetero colombiano Colombia, 1999, , Orientados: Carlos Mauricio Molina, Biología, Pontificia Universidad Javeriana - Puj - Sede Bogotá. Autores: LUZ STELLA BARRERO MENESES, CARLOS MAURICIO MOLINA,

3.- **Trabajo de conclusión de curso de pregrado** : Caracterización molecular y análisis de variabilidad genética en aislamientos de *Mycosphaerella fijiensis* provenientes de las regiones del Magdalena y Santander Colombia, 1999, , Orientados: Sergio Mauricio Aponte, Biología, Pontificia Universidad Javeriana - Puj - Sede Bogotá. Autores: LUZ STELLA BARRERO MENESES, SERGIO MAURICIO APONTE,

- 4.- **Iniciación Científica** : Caracterización molecular de lulo y tomate de árbol del banco de germoplasma de Corpoica mediante el empleo de marcadores COSII
Colombia, 2007, , Orientados: Felix Eugenio Enciso, Genética, Universidad Pedagógica Y Tecnológica De Colombia - Uptc - Sede Tunja.
Autores: LUZ STELLA BARRERO MENESES,
- 5.- **Tesis de doctorado** : Transposable elements donate lineage-specific regulatory sequences to host genomes
Estados Unidos, 2004, Tutor principal, Orientados: Kali C. Lewis, IRTA Fellow, National Institutes of Health.
Autores: LEONARDO MARINO-RAMIREZ, KALI C LEWIS,
- 6.- **Iniciación Científica** : Co-evolutionary Rates of Functionally Related Yeast Genes
Estados Unidos, 2005, , Orientados: Natalie Kantz, , National Institutes of Health.
Autores: LEONARDO MARINO-RAMIREZ, NATALIE KANTZ,
- 7.- **Iniciación Científica** : Ab initio prediction of Transcription Factor Binding Sites in the intergenic spacers of Histone Core Genes
Estados Unidos, 2007, , Orientados: Oscar Camilo Bedoya-Reina, , National Institutes of Health.
Autores: LEONARDO MARINO-RAMIREZ,
- 8.- **Iniciación Científica** : Statistical analysis of over-represented words in human promoter sequences
Estados Unidos, 2004, , Orientados: Gavin C. Kanga, IRTA Fellow, National Institutes of Health.
Autores: LEONARDO MARINO-RAMIREZ,

Demás trabajos

- 1.- **Demás trabajos** : Prediction of TFBS in intergenic spacers of histone core genes.
Colombia, 2007, , , , Presentación.
Autores: OSCAR CAMILO BEDOYA REINA, LEONARDO MARINO RAMIREZ,
- 2.- **Demás trabajos** : Taller Internacional Recursos Genéticos en el Trópico
Colombia, 2005, , , , Presentar el proyecto genoma internacional de la familia Solanaceae.
Autores: LUZ STELLA BARRERO MENESES, MARIO LOBO,
- 3.- **Demás trabajos** : Prediction of TFBS in intergenic spacers of histone core genes.
Estados Unidos, 2007, , , , Presentación.
Autores: LEONARDO MARINO-RAMIREZ, OSCAR CAMILO BEDOYA REINA,

Jurado/Comisiones evaluadoras de trabajo de grado

- 1.- **Curso de perfeccionamiento/especialización** : Identificación de nuevos genes CryI en aislamientos nativos de *Bacillus thuringiensis* por LSSP-PCR
Colombia, 2005, , Orientados: Alejandro Lopez, , Pontificia Universidad Javeriana - Puj - Sede Bogotá.
Autores: LUZ STELLA BARRERO MENESES,

Participación en comités de evaluación

- 1.- **Otra** : Par académico
Colombia, 2008, Ministerio De Agricultura Y Desarrollo Rural - Minagricultura.
Autores: SILVIA GOMEZ DAZA,
- 2.- **Otra** : Par académico
Colombia, 2007, Revista de agronomía Colombiana.
Autores: SILVIA GOMEZ DAZA,
- 3.- **Otra** : Par académico
Colombia, 2004, Agro-Bio.
Autores: SILVIA GOMEZ DAZA,
- 4.- **Otra** : Par académico
Colombia, 2005, Pontificia Universidad Javeriana - Puj - Sede Bogotá.
Autores: SILVIA GOMEZ DAZA,
- 5.- **Otra** : Par académico
Colombia, 2008, Ministerio de Agricultura y Desarrollo Rural.
Autores: SILVIA GOMEZ DAZA,
- 6.- **Otra** : Evaluador artículo revista Facultad Nacional de Agronomía, Medellín
Colombia, 2005, .
Autores: LUZ STELLA BARRERO MENESES,
- 7.- **Otra** : Evaluador revista Facultad Nacional de Agronomía, Medellín
Colombia, 2005, .

Autores: LUZ STELLA BARRERO MENESES,

8.- **Concurso público** : Evaluador proyectos y programas Colciencias
Colombia, 2007, colciencias.

Autores: LUZ STELLA BARRERO MENESES,

9.- **Concurso público** : Evaluador proyectos Colciencias
Colombia, 2008, .

Autores: LUZ STELLA BARRERO MENESES,

10.- **Concurso público** : Evaluador proyectos MADR

Colombia, 2005, Ministerio De Agricultura Y Desarrollo Rural - Minagricultura.

Autores: LUZ STELLA BARRERO MENESES,

Trabajos dirigidos/Tutorías en marcha

1.- **Iniciación Científica** : Construcción de mapas genéticos comparativos en las solanáceas lulo y tomate de árbol mediante el empleo de marcadores COSII

Colombia, 2006, , Orientados: Zulma Cardenas, , .

Autores: LUZ STELLA BARRERO MENESES, ZULMA CARDENAS,

2.- **Iniciación Científica** : Mapeo de genes candidatos involucrados en induccion de resistencia y crecimiento vegetal de tomate tratado con *Trichoderma koningii* (TH003)

Colombia, 2006, , Orientados: Ivan Fernando Calixto, , Universidad Pedagógica Y Tecnológica De Colombia - Uptc - Sede Tunja.

Autores: LUZ STELLA BARRERO MENESES, IVAN FERNANDO CALIXTO,

3.- **Iniciación Científica** : Caracterización molecular de la colección colombiana de lulo y tomate de árbol mediante el uso de marcadores COSII

Colombia, 2006, , Orientados: Felix Enciso, , Universidad Pedagógica Y Tecnológica De Colombia - Uptc - Sede Tunja.

Autores: LUZ STELLA BARRERO MENESES, FELIX ENCISO,

4.- **Iniciación Científica** : Evaluación de polimorfismos para la construcción de mapas genéticos comparativos en las solanáceas lulo y tomate de árbol mediante el empleo de marcadores COSII

Colombia, 2006, , Orientados: Zulma Cardenas, , .

Autores: LUZ STELLA BARRERO MENESES, ZULMA CARDENAS,

5.- **Iniciación Científica** : Evaluación de polimorfismos con marcadores COSII para la construcción de un mapa genético en papa criolla

Colombia, 2006, , Orientados: Ivan Fernando Calixto, , Universidad Pedagógica Y Tecnológica De Colombia - Uptc - Sede Tunja.

Autores: LUZ STELLA BARRERO MENESES, IVAN FERNANDO CALIXTO,

6.- **Tesis de maestría** : Caracterización morfo agronómica y molecular de la colección de mora (*Rubus glaucus* Bent) de CORPOICA y material de agricultor

Colombia, 2006, Tutor principal, Orientados: Natalia Espinoza, Quimica, Universidad Nacional De Colombia - Sede Bogotá.

Autores: LUZ STELLA BARRERO MENESES, NATALIA ESPINOZA,

7.- **Iniciación Científica** : Mapeo de genes relacionados con induccion de resistencia y crecimiento vegetal por *Trichoderma* en tomate

Colombia, 2007, , Orientados: Jaime Simbaqueba, Biotecnología, Universidad Militar Nueva Granada - Unimilitar.

Autores: LUZ STELLA BARRERO MENESES,

8.- **Iniciación Científica** : Co-evolutionary Rates of Functionally Related Yeast Genes

Estados Unidos, 2005, , Orientados: Natalie Kantz, , National Institutes of Health.

Autores: LEONARDO MARINO-RAMIREZ,

9.- **Tesis de doctorado** : Transposable elements donate lineage-specific regulatory sequences to host genomes

Estados Unidos, 2004, Tutor principal, Orientados: Kali C. Lewis, IRTA Fellow, National Institutes of Health.

Autores: LEONARDO MARINO-RAMIREZ,

10.- **Iniciación Científica** : Statistical analysis of over-represented words in human promoter sequences

Estados Unidos, 2004, , Orientados: Gavin C. Kanga, IRTA Fellow, National Institutes of Health.

Autores: LEONARDO MARINO-RAMIREZ,

11.- **Tesis de doctorado** : DNA Structure Conservation in Functional Elements

Estados Unidos, 2008, Coturor/asesor, Orientados: Loren Hansen, IRTA Fellow, National Institutes of Health.

Autores: LEONARDO MARINO-RAMIREZ,

Empresas de I+D

Proyectos

- 1.- Estimación de la diversidad genética de acciones de Uchuva con el uso de marcadores COSII; 2008 -
- 2.- Marcadores bioquímicos en un programa de introgresión rápida de genes entre las especies aceiteras *Elaeis guineensis* y *E. oleifera*; -
- 3.- Marcadores bioquímicos y moleculares en la certificación de semilla híbrida comercial de palma de aceite (*E. guineensis*); -
- 4.- Mejoramiento de variedades colombianas de algodón mediante la introducción de genes BT y RR mediante cruza y retrocruza con variedades transgénicas.; 2006 -
En Colombia se ha generado materiales genéticos de algodón con buenas características de producción, calidad y adaptación apropiada a nuestro ecosistema, sin embargo la producción comercial tiene como gran limitante la dificultad en el manejo de insectos plagas y malezas. A través de este proyecto se pretende mejorar las variedades Colombianas mediante la introducción de genes que transfieran resistencia a lepidópteros y tolerancia a glifosato a través de cruza y retrocruza con materiales transgénicos mejorando la sostenibilidad y competitividad del cultivo en el país. El seguimiento a las variedades colombianas de algodón que se le han introducido los genes Bt y RR se realizará en cada una de las retrocruzas a nivel de campo y laboratorio.
- 5.- Identificación y aislamiento de genes de resistencia a Sigatoka negra de germoplasma de babano; -
- 6.- Evaluación de la actividad insecticida de cepas nativas de *Bacillus thuringiensis* para el control biológico de plagas de importancia económica en el cultivo del algodón (Fase 2); 2001 - 2002
En esta segunda fase se realizó la caracterización microscópica, bioquímica y molecular de otras cepas nativas de *Bacillus thuringiensis*, y se realizó bioensayos para el control de *Spodoptera frugiperda* y *Alabama argillaceae*. Los resultados mostraron 2 cepas nativas como promisorias por su actividad tóxica a nivel de ensayos en laboratorio y en parcelas semicontroladas de cultivo de algodón.
- 7.- Detección Molecular de alelos R1 y R2 que confieren resistencia a *Phytophthora infestans* Mont, De Bary en genotipos tetraploides de papa.; 2001 - 2001
- 8.- Caracterización molecular de microorganismos asociados a procesos de nitrificación y desnitrificación de aguas de refinería; 1999 - 2001
- 9.- Transformación genética de papa con los genes *cry1Ab* y *cry1Ac* de *Bacillus thuringiensis* para el posible control de *Tecia solanivora*; 2000 -
En este proyecto se están transformando plantas de papa con los genes *cry1Ab* y *cry1Ac* de *Bacillus thuringiensis* con una posible resistencia a *Tecia solanivora*. La transformación genética se realiza usando *Agrobacterium*. Las líneas transformadas se evaluarán mediante análisis genéticos para determinar la integración y expresión de los genes transferidos. Los tubérculos de las líneas con mayor expresión serán evaluados en bioensayos para determinar la eficiencia en el control de *Tecia solanivora*.
- 10.- Determinación de niveles y calidad de almidones de la CCC de papa *Solanum phureja* para futuros proyectos de mejoramiento; 2001 -
- 11.- Identificación molecular de las variedades e híbridos generados por Corpoica.; 2001 - 2001
- 12.- Introducción de resistencia al picudo del algodón *Anthonomus grandis* a variedades comerciales colombianas mediante transformación genética; 2001 -
- 13.- Obtención de nuevas variedades de algodón para mejorar la competitividad de la producción. Transformación genética de variedades colombianas de algodón.; 2001 - 2001
- 14.- Obtención de variedades de *Musa* (Hartón Común) con resistencia al picudo negro *Cosmopolites sordidus* mediante transformación genética.; 2001 -
- 15.- Uso de técnicas de microinyección para el cultivo de cítricos en Colombia.; 2001 - 2001
- 16.- Cultivo de Tejidos para la Propagación masiva y transformación genética de guayaba para la producción de fruta sin semilla y larga vida.; 2001 -
- 17.- Establecimiento de un método de transformación genética de *Trichoderma koningii* Th003 con un gen reportero; 2007 -
Estandarizar un protocolo de transformación genética de *Trichoderma koningii* utilizando el gen reportero GFP para usarlo como base de futuros trabajos.
- 18.- Caracterización molecular de los hongos entomopatógenos *Beauveria bassiana* y *Metarhizium anisopliae*; 1996 - 1998
En este proyecto se utilizaron las técnicas de RAPD y amplificación de ITS para la caracterización molecular de los hongos entomopatógenos *Beauveria bassiana* y *Metarhizium anisopliae*. Este proyecto hacía parte de un macroproyecto que buscaba el mejoramiento genético de estos hongos entomopatógenos.
- 19.- Biorremediación de fenol en aguas de la industria petrolera; 1999 - 2001
Con este proyecto se aislaron y caracterizaron bacterias con potencial biológico en la remoción de compuestos fenólicos provenientes de agua de refinería.
- 20.- Caracterización biológica y molecular de cepas de *Bacillus thuringiensis* para el control de insectos plagas en

agricultura; 1997 - 1999

Las pérdidas de los cultivos, debido al ataque de insectos van desde el 10 al 30% de los costos totales de producción. Generalmente las plagas se han venido controlando con insecticidas químicos que traen serios problemas tanto para la salud humana como para los ecosistemas. Una alternativa de bajo impacto ecológico para el control de plagas es el uso de entomopatógenos como *Bacillus thuringiensis*. Dada la diversidad existente en Colombia y el potencial de encontrar microorganismos entomopatógenos mejores adaptados a las condiciones del trópico y con mayor especificidad contra los insectos plagas nativos; nace este proyecto tendiente al aislamiento y caracterización microscópica, bioquímica y molecular de cepas nativas de *B. thuringiensis* tomadas del Banco de Cepas de Corpoica. La información obtenida permitió clasificar aproximadamente 680 aislamientos de acuerdo con su posible actividad biológica contra insectos. También dentro de este proyecto se aisló, expresó y purificó la proteína Cry 3Aa recombinante de *Bacillus thuringiensis* var. *san diego* así como se realizó una evaluación biológica sobre larvas de primer instar de *Premnotrypes vorax* Hustache con el fin de que en trabajos futuros se pudiera incorporar el gen cry 3Aa en plantas mediante técnicas de ingeniería genética.

21.- Evaluar la actividad insecticida de cepas nativas de *Bacillus thuringiensis* sobre plagas de importancia económica en el cultivo de algodón; 2000 - 2001

Una alternativa de menor impacto ecológico para el control de plagas es el uso de microorganismos entomopatógenos con *Bacillus thuringiensis*. Aprovechando la gran biodiversidad de nuestro país este proyecto fue encaminado a la caracterización microscópica, bioquímica y molecular de cepas nativas de *B. thuringiensis* que se encuentran mejor adaptadas a nuestras condiciones ambientales. Con estas cepas se realizaron bioensayos a nivel de laboratorio para el control de *Spodoptera frugiperda* y *Alabama argillaceus* que son plagas que afectan los cultivos de maíz y algodón generando altos costos en la producción y disminución en la productividad de dichos cultivos.

22.- Characterization and gene function of *zrp4* gene, a root preferential gene of corn; 1993 - 1998

23.- Developing of maize anther culture method for genetic improvement; 1991 - 1994

24.- Production of male sterile plants in corn by genetic transformation; 1998 - 2000

25.- characterization of transgenic corn plants with altered starch quality; 1998 - 2000

26.- Desarrollo de una metodología para diagnóstico de sexos en papaya; 2001 -

27.- Base molecular de forma y tamaño del fruto del tomate: Un modelo para variación de características cuantitativas; 2003 - 2006

La forma y el tamaño del fruto son factores importantes que determinan la producción, la calidad y la aceptación del consumidor en muchos cultivos. Ambas son características de herencia cuantitativa que han sido difíciles de entender con herramientas de biología molecular. El proyecto busca identificar, aislar y entender la base molecular de loci (sitios del genoma) que afectan el tamaño y la forma del fruto del tomate. Los resultados contribuirán a vislumbrar la naturaleza de la base molecular de la variación de características cuantitativas y contribuirán a entender la transformación de los ovarios -pequeños órganos reproductivos- a frutos grandes de diferentes formas y tamaños que asociamos con la agricultura moderna. Además, se reconstruirán los eventos involucrados en la domesticación del tomate y otras especies portadoras de frutos.

28.- Development of COS databases for the Andean fruited species lulo and tree tomato-DCC Plant Genome Research- NSF Supplement Plant Genomics Project: "Sequence and annotation of the tomato euchromatin: A framework for Solanaceae Comparative Biology; 2005 - 2007

Dos especies exóticas nativas de la región andina (Colombia, Ecuador, y Perú), el lulo (*Solanum quitoense*) y el tomate de árbol (*Solanum betaceum*) tienen un gran potencial para convertirse en productos de alto impacto en mercados nacionales e internacionales con un alto retorno económico para los agricultores. A pesar de su incrementado valor en el mercado, los mayores limitantes para su adopción por agricultores locales son la falta de sustento tecnológico asociado con una oferta casi nula de materiales mejorados. CORPOICA mantiene colecciones de lulo y tomate de árbol y taxa relacionados, las cuales tienen procesos parciales de caracterización fenotípica y genotípica. El desarrollo de marcadores moleculares informativos para la caracterización de estos recursos genéticos es una necesidad inmediata para su apropiado uso. Estos marcadores serán también útiles en programas de introgresión de alelos para resistencia, alta producción y calidad. El proyecto propone el desarrollo y entrenamiento de científicos colombianos en la generación de bases de datos de secuencias parciales de genes COSII (segunda generación de secuencias de genes ortólogos conservados) en lulo y tomate de árbol, de tal forma que esta información pueda ser usada a futuro para evaluar la diversidad funcional en colecciones de germoplasma, desarrollar mapas genéticos y asociativos, y asistir programas de mejoramiento genético a través de selección asistida por marcadores y genómica.

29.- Análisis de la expresión génica de plantas de tomate tratadas con *Trichoderma koningii* (TH003) en relación con parámetros de inducción de la resistencia y del crecimiento vegetal; 2006 - 2008

Se pretende estimular a las comunidades científica y académica para que aborden campos del conocimiento tales como la resistencia inducida, la estimulación del crecimiento vegetal y la genómica estructural y funcional como elementos de la biotecnología actual indispensables para los estudios de bioprospección y de elucidación de fenómenos biológicos. Igualmente se pretende continuar con la sensibilización de los agricultores sobre los beneficios directos e indirectos del control biológico de fitopatógenos y del uso de estrategias de producción limpias

30.- Certificación y escalamiento de material de mora con potencial nutritivo y nutraceutico para entrega a pequeños agricultores; 2006 - 2009

Dentro de la cadena frutícola, la mora (*Rubus glaucus* Bent) es un frutal andino que representa una de las 10 frutas agroindustriales promisorias. En Colombia se cultiva principalmente la mora de Castilla, variedad ampliamente adaptada

pero que presenta limitaciones de susceptibilidad fitosanitaria y bajo contenido de grados Brix. La oferta de un solo material puede conducir a la vulnerabilidad en la producción. Esto aunado a la necesidad cada vez más creciente de producir materiales con un alto valor nutritivo, hacen necesario tomar acciones que conlleven a seleccionar materiales con valor agregado que contribuyan a mejorar la calidad de vida. El proyecto busca fortalecer la cadena de la mora e impulsar su crecimiento, a través del desarrollo tecnológico integrado, que involucre diferentes líneas de producción. Para esto, se trabaja en la caracterización e identificación de material promisorio (30 accesiones de la colección CORPOICA y materiales de productores de Cundinamarca) desde el punto de vista morfo-agronómico, nutricional y nutraceutico con actividad antioxidante. Se utilizan marcadores moleculares tipo AFLPs para estudios de huella genómica. El material élite seleccionado es utilizado para el desarrollo de esquemas de producción limpia de semilla que involucren la utilización de biocontroladores y biofertilizantes que mejoren su establecimiento y reduzcan los problemas fitosanitarios originados durante la propagación. El material élite es escalado mediante multiplicación masiva in vitro para su entrega a pequeños agricultores quienes derivan su sustento del cultivo de la mora.

31.- Iniciativa para la aplicación de la genómica en el mejoramiento genético de la papa criolla: estudio de la resistencia a la gota (*Phytophthora infestans*); 2006 - 2008

La papa (*Solanum* sp) es el cuarto cultivo de importancia alimentaria en el mundo después del maíz, el trigo y el arroz. La papa criolla (*S. phureja*) como alimento, ofrece a la dieta del consumidor un excelente valor nutricional. Como recurso genético, ofrece al país una gran variabilidad genética para la generación de variedades con características de interés agronómico entre las cuales se destaca la resistencia genética a plagas y enfermedades. La papa es atacada por una gran número de patógenos, siendo el principal *Phytophthora infestans*, causante del tizón tardío o gota. En Colombia esta enfermedad es considerada como un problema altamente prioritario, que puede llegar a causar hasta el 100% de pérdida cuando el cultivo no es adecuadamente protegido. Entre las estrategias de control de *P. infestans*, el método más eficiente sigue siendo la utilización de fungicidas y, el más deseado, apropiado y realista, la utilización de variedades resistentes. El presente proyecto propone utilizar 2 parentales de *S. phureja* (uno altamente resistente y otro altamente susceptible a *P. infestans*) y su población F1 para ser caracterizados molecularmente a través del aislamiento de genes de defensa que se expresan como resultado de la infección y del uso de marcadores SSRs y COSII para la futura selección asistida por marcadores y/o genómica en el mejoramiento genético de papa diploide (*S. phureja*) y papa tetraploide (*S. tuberosum*).

32.- Uso de marcadores COS en las solanáceas lulo y tomate de árbol para caracterización de germoplasma y genómica comparativa; 2006 - 2009

Los frutales andinos de la familia Solanácea tales como el lulo (*Solanum quitoense*) y el tomate de árbol (*Cyphomandra betacea*) tienen un gran potencial para convertirse en productos de alto impacto en mercados nacionales e internacionales con un alto retorno social y económico a agricultores y consumidores. A pesar de su incrementado valor en el mercado, los principales limitantes para su adopción por agricultores locales son la falta de sustento tecnológico asociado con una oferta casi nula de materiales mejorados. En este sentido el presente proyecto pretende contribuir al conocimiento de la base genética de la diversidad de estos frutales como insumo para programas de mejoramiento en Colombia. Se están utilizando marcadores COSII (segunda generación de secuencias de genes ortólogos conservados), los cuales se generan a partir de secuencias conservadas de genes de especies de la familia Solanácea. Los marcadores COS están siendo utilizados en procesos de caracterización de germoplasma, identificación de híbridos, y estudios iniciales de genómica comparativa con otras especies de la familia Solanácea para contribuir al futuro mejoramiento asistido por marcadores o por genómica en estos cultivos.

33.- Evaluación de resistencia de uchuva a *F. oxysporum* como control integrado de la enfermedad; 2008 - 2011

Dentro de la cadena frutícola, la uchuva (*Physalis peruviana* L.) representa una de las especies con mayor proyección para emprender proyectos productivos con miras a la exportación. La uchuva es la segunda fruta de exportación para Colombia después del banano. El genotipo colombiano es el más apetecido por su sabor más dulce y por su mejor color, confiriendo así una ventaja en los mercados internacionales. Sin embargo, a pesar de su riqueza y gran potencial, la uchuva no ha adquirido el grado de importancia esperado, lo cual puede atribuirse a una falta de sustento tecnológico adecuado. Su cultivo en Colombia enfrenta importantes problemas fitosanitarios, dentro de los cuales se destaca el marchitamiento causado por el hongo *Fusarium oxysporum* que está afectando seriamente la sostenibilidad económica y ambiental del cultivo, debido a la magnitud de las pérdidas que reporta el sector productivo y al abuso de fungicidas de síntesis química que se emplean para su control. En el mercado mundial la tendencia actual de la demanda de alimentos se caracteriza por el consumo de productos "limpios". Al respecto, la legislación en varios países exige una disminución considerable del uso de plaguicidas químicos, la implementación de medidas fitosanitarias integradas y la utilización de buenas prácticas agrícolas (<http://www.cci.org.co/noticias.html>; <http://www.eurep.org/fruit/index.html>). En este contexto, la resistencia varietal junto con otros métodos de control integrado (como el control biológico, ver proyecto 2 del presente programa) constituye una herramienta valiosa para reducir los riesgos de contaminación ambiental y de residualidad en el producto de consumo, provocados por los ingredientes activos utilizados habitualmente. La utilización del potencial genético de cualquier cultivo, depende de la disponibilidad de una base genética amplia para poder aplicar procesos de selección. En uchuva, las instituciones portadoras del germoplasma (Universidad Nacional, Universidad de Nariño, Corpoica) han realizado estudios morfo-agronómicos y de calidad del fruto. Sin embargo, no se ha estudiado el potencial de las colecciones para atributos de resistencia a enfermedades limitantes, lo cual constituye el primer paso para la generación de variedades y material de siembra con tolerancia o resistencia. El patógeno más limitante en la actualidad es *F. oxysporum*, el cual debido a su capacidad de reproducción, diseminación y formación de estructuras de resistencia, presenta una larga persistencia en los suelos y gran resistencia a los métodos convencionales de control. Esto hace que la utilización de variedades resistentes o tolerantes sea el método de control deseable. El presente proyecto propone complementar el programa focalizado en el patógeno, con un enfoque dirigido hacia la planta. Para ello, se evaluará la

variabilidad genética de la uchuva y se identificarán fuentes de resistencia/tolerancia a *F. oxysporum*. Paralelamente, se buscarán homólogos de genes de resistencia en uchuva que puedan estar asociados con resistencia a dicho patógeno, para lo cual será necesario conocer la estructura poblacional con marcadores altamente informativos. El estudio aportará herramientas valiosas para la identificación de materiales con tolerancia o resistencia, el diseño de estrategias de cruzamiento y la generación eficiente de variedades resistentes. De esta forma se espera que los resultados del proyecto contribuyan a impactar la producción limpia de uchuva a través de esquemas de manejo integrado de la enfermedad más limitante, y así, la competitividad y sostenibilidad de la cadena en los mercados nacionales e internacionales.

34.- Análisis bioquímico de la interacción huesped - patógeno y desarrollo de un método para el diagnóstico de la enfermedad del anillo rojo en palma de aceite; -

35.- Genes de defensa a Sigatoka en America Latina; -

36.- Introduccion de genes con propiedades de resistencia al *Anthonomus grandis* Boheman a variedades Colombianas comerciales de algodón mediante transformacion genetica; 2003 - 2004

Anthonomus grandis es una plaga que afecta el cultivo de algodón debido a que causa daño a las estructuras productivas (botones florales y cápsulas), esta plaga es controlada con químicos que causan daño al hombre, medio ambiente e insectos benéficos; una alternativa de menos impacto ecológico es la obtención de plantas que se defiendan del ataque de los insectos disminuyendo además los costos de producción. Por lo anterior este proyecto fue encaminado en dos vías: una en la búsqueda de genes que expresen proteínas tóxicas a *A. grandis* las cuales fueron probadas a través de bioensayos a nivel de laboratorio y se le determinó su toxicidad sobre larvas de primer instar, para luego estos genes ser introducidos en materiales de algodón y la otra vía fue estandarizar un sistema de regeneración in Vitro de variedades mejoradas de algodón a través de embriogénesis y/o organogénesis.

37.- Desarrollo de una técnica eficiente de regeneración in vitro en variedades comerciales Colombianas de algodón para introducir genes para el control de picudo; 2004 - 2004

La obtención de plantas transgénicas se puede lograr a través de dos vías: una directa mediante la introducción de los genes de interés utilizando *Agrobacterium tumefaciens* y/o biobalística y la otra indirecta por cruza y retrocruza entre los padres contranste para la característica deseada. La primera vía fue la escogida en este proyecto, para introducir genes que confieran resistencia a las variedades colombianas de algodón de ataque de *Anthonomus grandis* (picudo del algodón). Para poder lograr el objetivo se trabajó en la estandarización de un protocolo de transformación genética utilizando *A. tumefaciens* y en un sistema de regeneración con las variedades de algodón.

38.- Aislamiento y caracterización de los genes nag70 y gluc78 de *Trichoderma koningii* cepa th003; 2007 -

Utilizando las regiones conservadas entre las especies del mismo género a través de PCR específica se pretende amplificar los genes Nag 70 y Gluc 78 de *Trichoderma Koningii* involucrados en la actividad micoparasítica, para luego clonarlos y caracterizarlos.



Departamento Administrativo de
Ciencia, Tecnología e Innovación
Colciencias
República de Colombia

PUBLICACIÓN DE RESULTADOS

CONVOCATORIA NACIONAL PARA MEDICIÓN DE GRUPOS DE INVESTIGACIÓN EN CIENCIA, TECNOLOGÍA E INNOVACIÓN AÑO 2010 – No. 509 – 2010

*Se anexa el listado ordenado de manera ascendente
por código de identificación del grupo de investigación*

**CONVOCATORIA NACIONAL PARA MEDICIÓN DE GRUPOS DE INVESTIGACIÓN EN CIENCIA, TECNOLOGÍA E INNOVACIÓN
AÑO 2010 – No. 509 – 2010**

- Identificar nuevos grupos de investigación que cumplen con los requisitos exigidos para considerarse como tales.
- Clasificar los grupos de investigación científica, tecnológica o de innovación colombianos de acuerdo a su producción.
- Contar con información actualizada para generar estadísticas precisas y confiables sobre las capacidades del Sistema Nacional de Ciencia y Tecnología del país.
- Identificar características y condiciones de los grupos de investigación que permitan posibles reagrupaciones y comparaciones más avanzadas.

TOTAL DE GRUPOS : 4.072

CÓDIGO DEL GRUPO	NOMBRE DEL GRUPO DE INVESTIGACIÓN	ENTIDADES A LAS QUE PERTENECE EL GRUPO	CATEGORÍA
COL0076853	AGRICULTURA SOSTENIBLE	UNIVERSIDAD DE CORDOBA - UNICOR	C
COL0076879	ESTUDIOS LITERARIOS Y CULTURALES	UNIVERSIDAD SERGIO ARBOLEDA	C
COL0077009	IDENTIDADES	UNIVERSIDAD DEL PACÍFICO	C
COL0077019	CES CULTURA, EDUCACION Y SOCIEDAD	CORPORACIÓN UNIVERSITARIA DE LA COSTA - CUC	B
COL0077020	DESARROLLO REGIONAL - IDER	UNIVERSIDAD DE NARIÑO	C
COL0077055	DERECHO PRIVADO	UNIVERSIDAD LIBRE DE COLOMBIA - CARTAGENA	C
COL0077064	INGENIERIA DEL SOFTWARE Y REDES	CORPORACIÓN UNIVERSITARIA DE LA COSTA - CUC	B
COL0077091	GRUPO DE INVESTIGACIÓN EN SOCIEDAD, EMPRESA Y MEDIO AMBIENTE (GISEMA)	UNIVERSIDAD LIBRE DE COLOMBIA - CARTAGENA	C
COL0077126	GRUPO DE INVESTIGACIÓN EN CULTURA Y ESPIRITUALIDAD (INCE)	UNIVERSIDAD PONTIFICIA BOLIVARIANA - SEDE MEOELLÍN	D
COL0077153	POLÍTICAS DE DESARROLLO	PONTIFICIA UNIVERSIDAD JAVERIANA	D
COL0077189	LABORATORIO DE INVESTIGACIONES FITODUIMICAS Y FARMACOLOGICAS DE LA UNIVERSIDAD DE CARTAGENA (LIFFUC)	UNIVERSIDAD DE CARTAGENA - UNICARTAGENA; E.S.E. HOSPITAL UNIVERSITARIO DEL CARIBE	D
COL0077206	DRISO - DESARROLLO E INGENIERIA DE SOFTWARE	UNIVERSIDAD ICESI	D
COL0077251	GESTION INTEGRAL DE LAS ORGANIZACIONES	UNIVERSIDAD ANTONIO NARIÑO	D
COL0077378	PRODUCIENDO	UNIVERSIDAD DE LA AMAZONIA	D
COL0077387	GRUPO DE INVESTIGACIÓN PARA EL DESARROLLO, LA PAZ Y LA DEMOCRACIA	UNIVERSIDAD DE SAN BUENAVENTURA	C
COL0077439	GRUPO DE INVESTIGACIÓN ODONTOLÓGICO FORENSE	UNIVERSIDAD METROPOLITANA - UMET	D
COL0077565	METROLOGIA	UNIVERSIDAD TECNOLÓGICA DE PEREIRA - UTP	C
COL0077583	GRUPO PRODUCTOS VERDES (GPV)	UNIVERSIDAD DE PAMPLONA - UDP	D
COL0077592	GRUPO DE INVESTIGACION EN PSICOLOGIA GIPSI	UNIVERSIDAD DE SAN BUENAVENTURA	D
COL0077707	FITOPATOLOGÍA Y ECOFISIOLOGÍA VEGETAL	UNIVERSIDAD MILITAR NUEVA GRANADA - UNIMILITAR	D
COL0077725	LALETUS (LANGUAGE LEARNING AND TEACHING - UNIVERSIDAD DE LA SABANA)	UNIVERSIDAD DE LA SABANA - UNISABANA	D
COL0077734	CULTURA Y EDUCACIÓN	UNIVERSIDAD DE PAMPLONA - UDP	D
COL0077743	GRUPO DE INVESTIGACIÓN DE ARQUEOLOGÍA, PATRIMONIO Y AMBIENTE REGIONALES - ARQUEO.REGIÓN	UNIVERSIDAD DEL TOLIMA	B
COL0077789	MUNDOS VIRTUALES	UNIVERSIDAD NACIONAL DE COLOMBIA	D
COL0077799	INVESTIGACIÓN-IT UNIMINUTO	CORPORACIÓN UNIVERSITARIA MINUTO DE DIOS	C
COL0077814	AGROBIOLOGÍA DE ESPECIES VEGETALES PROMISORIAS DE CLIMA FRÍO	UNIVERSIDAD MILITAR NUEVA GRANADA - UNIMILITAR	D
COL0077859	PROSPECTIVA ESTRATÉGICA ORGANIZACIONAL - PESTO	PROSERES PROSPECTIVA ESTRATÉGICA LTDA.	D
COL0077877	IPPE (INVESTIGACIÓN EN PRÁCTICAS PEDAGÓGICAS)	UNIVERSIDAD SURCOLOMBIANA - USCO	D
COL0078025	INTEROPERABILIDAD TECNOLÓGICA Y SEMANTICA	UNIVERSIDAD DISTRITAL "FRANCISCO JOSÉ DE CALDAS"	D
COL0078079	CALIDAD DE VIDA Y POLÍTICA SOCIAL	UNIVERSIDAD DE LA SALLE - UNISALLE	D
COL0078099	GIRA, GRUPO DE INVESTIGACIÓN EN REFRIGERACIÓN Y AIRE ACONDICIONADO	FUNDACION UNIVERSIDAD DE AMERICA	D
COL0078123	CONOCIMIENTO, MANEJO Y CONSERVACIÓN DE LOS ECOSISTEMAS DEL CHOCÓ BIOGEOGRÁFICO	INSTITUTO DE INVESTIGACIONES AMBIENTALES DEL PACÍFICO "JHON VON NEUMANN"	D
COL0078203	ESTUDIOS AMBIENTALES Y DEL HÁBITAT	UNIVERSIDAD DEL TOLIMA	D
COL0078221	ESPACIOS FUNCIONALES	UNIVERSIDAD DEL CAUCA - UNICAUCA	D
COL0078277	GRUPO DE INVESTIGACIÓN Y DESARROLLO EN ORGANIZACIONES, SISTEMAS Y COMPUTACIÓN	UNIVERSIDAD DEL MAGDALENA - UNIMAGDALENA	D
COL0078286	GESTIÓN Y POLÍTICAS PÚBLICAS TERRITORIALES -GPPT-	UNIVERSIDAD NACIONAL DE COLOMBIA	D
COL0078295	SIGMA	PONTIFICIA UNIVERSIDAD JAVERIANA - PUJ - SEDE CALI	D
COL0078366	JUEGO, CUERPO Y MOTRICIDAD	UNIVERSIDAD DE LOS LLANOS - UNILLANOS	D
COL0078384	DERECHO, SOCIEDAD Y DESARROLLO	CORPORACION UNIVERSITARIA REPUBLICANA	C
COL0078428	MICOBAC-UN	UNIVERSIDAD NACIONAL DE COLOMBIA	A1
COL0078455	G-3IN	CORPORACIÓN UNIVERSITARIA LASALLISTA	D
COL0078464	PEDAGOGÍA, TECNOLOGÍA Y SOCIEDAD EN LAS ARTES VISUALES	PONTIFICIA UNIVERSIDAD JAVERIANA	D
COL0078599	GRUPO DE INTEGRIDAD ESTRUCTURAL GIE	UNIVERSIDAD DE LOS ANDES - UNIANDES	B

89 05

**CONVOCATORIA NACIONAL PARA MEDICIÓN DE GRUPOS DE INVESTIGACIÓN EN CIENCIA, TECNOLOGÍA E INNOVACIÓN
AÑO 2010 – No. 509 – 2010**

- Identificar nuevos grupos de investigación que cumplen con los requisitos exigidos para considerarse como tales.
- Clasificar los grupos de investigación científica, tecnológica o de innovación colombianos de acuerdo a su producción.
- Contar con Información actualizada para generar estadísticas precisas y confiables sobre las capacidades del Sistema Nacional de Ciencia y Tecnología del país.
- Identificar características y condiciones de los grupos de investigación que permitan posibles reagrupaciones y comparaciones más avanzadas.

TOTAL DE GRUPOS : 4.072

CÓDIGO DEL GRUPO	NOMBRE DEL GRUPO DE INVESTIGACIÓN	ENTIDADES A LAS QUE PERTENECE EL GRUPO	CATEGORÍA
COL0083212	GRUPO DE INVESTIGACIÓN EN DERECHO PRIVADO DE LA FACULTAD DE JURISPRUDENCIA DE LA UNIVERSIDAD DEL ROSARIO	UNIVERSIDAD DEL ROSARIO	B
COL0083259	MANEJO Y CONSERVACIÓN DE ECOSISTEMAS Y VIDA SILVESTRE	PROYECTO DE CONSERVACIÓN DE AGUAS Y TIERRAS	C
COL0083277	ESTUDIOS AMBIENTALES Y DEL TERRITORIO	UNIVERSIDAD LIBRE DE COLOMBIA - PEREIRA	D
COL0083319	SISTEMA ESTOMATOGNÁTICO Y MORFOFISIOLOGÍA	UNIVERSIDAD SANTO TOMÁS DE AQUINO - SEDE BUCARAMANGA - USTABUC	A
COL0083455	GRUPO DE INVESTIGACIÓN EN RELACIONES FAMILIARES	CORPORACIÓN UNIVERSITARIA ADVENTISTA	D
COL0083464	EDUCACIÓN SUPERIOR, CONOCIMIENTO Y GLOBALIZACIÓN	UNIVERSIDAD PEDAGÓGICA NACIONAL - U.P.N.	C
COL0083517	GRUPO DE INVESTIGACIÓN EN FILOSOFÍA - REMATA	UNIVERSIDAD DEL NORTE - UNINORTE	C
COL0083535	FILOSOFÍA Y ESCEPTICISMO	UNIVERSIDAD TECNOLÓGICA DE PEREIRA - UTP	C
COL0083698	INVESTIGACIÓN ANESTESIA CES - IACES	UNIVERSIDAD CES	D
COL0083731	PROCESOS Y MEDIOS DE COMUNICACIÓN	PONTIFICIA UNIVERSIDAD JAVERIANA - PUJ - SEDE CALI	D
COL0083749	GRUPO DE INVESTIGACIÓN EN MATERIALES, CATALISIS Y MEDIO AMBIENTE	UNIVERSIDAD NACIONAL DE COLOMBIA	C
COL0083849	MICROBIOLOGÍA Y MUTAGÉNESIS AMBIENTAL	UNIVERSIDAD INDUSTRIAL DE SANTANDER - UIS	C
COL0083867	TEMISA (TEOLOGÍA, ÉTICA Y MISIÓN SAN ALFONSO)	FUNDACIÓN UNIVERSITARIA SAN ALFONSO	D
COL0083965	EPIUIDES	UNIVERSIDAD DE SANTANDER	D
COL0083983	COHERENCIAS ENTRECRUZADAS	FUNDACIÓN UNIVERSITARIA SAN ALFONSO	D
COL0084013	ESTUDIOS SOCIOHISTÓRICOS DE LA SALUD Y LA PROTECCIÓN SOCIAL	UNIVERSIDAD NACIONAL DE COLOMBIA	C
COL0084059	ECONOMÍA, HISTORIA Y CONFLICTO	UNIVERSIDAD DE CARTAGENA - UNICARTAGENA	D
COL0084111	CÁLCULO CIENTÍFICO Y MODELAMIENTO MATEMÁTICO	UNIVERSIDAD NACIONAL DE COLOMBIA	D
COL0084176	CENTRO DE DESARROLLO INDUSTRIAL TECNOL	INDUSTRIAS TECNOLÓGICAS LIMITADA	D
COL0084209	GRUPO DE INVESTIGACIÓN EN ESTUDIOS AEROSPAZIALES	ESCUELA MILITAR DE AVIACIÓN "MARCO FIDEL SUÁREZ"	D
COL0084219	MODELAMIENTO PARA LA GESTIÓN DE OPERACIONES	UNIVERSIDAD NACIONAL DE COLOMBIA	D
COL0084247	GESTAR	FUNDACIÓN UNIVERSITARIA DEL ÁREA ANDINA	D
COL0084258	MEDICINA DEL ADULTO	UNIVERSIDAD DE LA SABANA - UNISABANA	C
COL0084318	DERECHOS COLECTIVOS Y AMBIENTALES-GIDCA	UNIVERSIDAD NACIONAL DE COLOMBIA	D
COL0084416	CRECIMIENTO, CAMBIO ESTRUCTURAL Y ECONOMÍA REGIONAL	FUNDACIÓN UNIVERSIDAD CENTRAL	B
COL0084505	CU	FUNDACIÓN LOGYCA	D
COL0084532	GRUPO DE INVESTIGACIÓN EN SALUD PÚBLICA	CORPORACIÓN UNIVERSITARIA IBEROAMERICANA (ANTES TECNOLÓGICO INPI)	D
COL0084569	GAOPE	UNIVERSIDAD TECNOLÓGICA DE PEREIRA - UTP	A
COL0084588	GHEMA	UNIVERSIDAD DE LOS LLANOS - UNILLANOS	D
COL0084639	PSICOLOGÍA Y CIUDADANÍAS INCLUYENTES	UNIVERSIDAD NACIONAL DE COLOMBIA; UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA; UNIVERSIDAD COOPERATIVA DE COLOMBIA	D
COL0084775	ALCOM	UNIVERSIDAD INDUSTRIAL DE SANTANDER - UIS	D
COL0084848	PLUDEHUCO	UNIVERSIDAD SANTO TOMÁS	D
COL0084826	VISION SANA: GRUPO DE OFTALMOLOGÍA DE ESTUDIO EN CORNEA, CATARATA Y CIRUGÍA REFRACTIVA	CLÍNICA DE OFTALMOLOGÍA DE CALI S.A.	C
COL0084889	SISTEMAS EN OPERACIONES Y DESARROLLO APLICADO	UNIVERSIDAD COOPERATIVA DE COLOMBIA	C
COL0084899	DETECCIÓN DE CONTAMINANTES Y REMEDIACIÓN (DECOR)	UNIVERSIDAD PONTIFICIA BOLIVARIANA SECCIONAL BUCARAMANGA	B
COL0085057	COMUNIDAD LÚDICA DE ESTUDIOS INTERDISCIPLINARIOS	FUNDACIÓN UNIVERSITARIA TECNOLÓGICO COMFENALCO CARTAGENA	D
COL0085075	ESTEL DINA	DIRECCIÓN NACIONAL DE ESCUELAS - POLICÍA NACIONAL DE COLOMBIA	D
COL0085262	GRUPO DE INVESTIGACIÓN EN SALUD COMUNITARIA	UNIVERSIDAD DE LA SABANA - UNISABANA	D
COL0085271	FINANCE - GRUPO DE INVESTIGACIÓN EN ECONOMÍA Y FINANZAS	INSTITUTO TECNOLÓGICO METROPOLITANO DE MEDELLÍN - I.T.M.	D
COL0085404	LOGICIEL	FUNDACIÓN UNIVERSITARIA DE POPAYÁN	D
COL0085459	GENÉTICA MOLECULAR VEGETAL, BIOLOGÍA COMPUTACIONAL Y BIOINFORMÁTICA	CORPORACIÓN COLOMBIANA DE INVESTIGACIÓN AGROPECUARIA - CORPOICA	A
COL0085478	ECCO - EMOCIÓN, COGNICIÓN Y COMPORTAMIENTO	UNIVERSIDAD PONTIFICIA BOLIVARIANA - SEDE MEDELLÍN	D

92 05

Formación de Recurso Humano

1. Docente en Programas de Doctorado

Universidad de La Sabana - Doctorado en Biociencias

<http://www.unisabana.edu.co/postgrados/doctorado-en-biociencias/nuestro-programa/>

Contacto:

Ing. Esperanza Carvajal Hoyos
Coordinadora Doctorado Biociencias
Facultad de Ingeniería
Universidad de La Sabana
doctorado.biociencias@unisabana.edu.co
Tel. 8615555 Ext. 2557
Cel. 3112376878

2. Docente en Programas de Doctorado

Universidad Nacional de Colombia - Doctorado en Ciencias Biomédicas

<http://www.medicina.unal.edu.co/doctcienciabiom/>

Contacto:

Carlos Alberto Parra López
Coordinador Doctorado en Ciencias Biomédicas
Facultad de Medicina
Universidad Nacional de Colombia
Tel. 316-5000 Ext. 15039 / 15016
e-mail: docbiomed_fmbog@unal.edu.co

3. Docente en Programas de Doctorado

Boston University – Bioinformatics

<http://www.bu.edu/bioinformatics/>

Contacto:

Tom Tullius
Life Science and Engineering Building (LSE)
24 Cummington St
Boston, MA

tullius@bu.edu
Tel: +1 617-353-2482

4. Docente en Programas de Maestría

Universidad de Los Andes – Ingeniería de Sistemas

<http://sistemas.uniandes.edu.co/sitio/maestria>

Contacto:

Silvia Takahashi
Profesor Asociado
Universidad de los Andes
Carrera 1 este N° 19A-40 of. ML 705
Bogotá, Colombia
stakahas@uniandes.edu.co
Tel: 3394949 Ext. 2870

4. Docente en Programas de Maestría

Johns Hopkins University – Bioinformatics

<http://advanced.jhu.edu/academic/biotechnology/ms-in-bioinformatics/>

Contacto:

Beth M. Lemkelde
Advanced Academic Programs
Krieger School of Arts and Sciences
Johns Hopkins University
1717 Massachusetts Avenue, NW
Washington, DC 20036
Beth_Lemkelde@jhu.edu
Tel. +1 202-452-1916



CURSO INTERNACIONAL DE BIOINFORMÁTICA: MANEJO DE LAS HERRAMIENTAS BÁSICAS

Octubre 11–14 de 2005



PRESENTACIÓN

A pesar de ser la rama de las ciencias biológicas más recientemente surgida, la bioinformática es el área del conocimiento que ha hecho los aportes más valiosos a muchos de los grandes logros científicos de nuestro tiempo; tal es el caso del proyecto genoma humano, el cual sin duda alguna ha revolucionado la forma de ver al hombre y ha ampliado las posibilidades de las ciencias. Gracias a la aplicación de la informática a las ciencias biológicas algunas labores como la creación de bancos de información, el procesamiento y análisis de resultados experimentales y la comparación de grandes cantidades de datos son cada vez más rápidas y confiables. En la actualidad buena parte del avance de los proyectos de investigación en biología molecular, bioquímica, genética y biotecnología depende del uso de las herramientas que provee la bioinformática.

JUSTIFICACIÓN

El inmenso caudal de información científica que se genera día a día, la profundidad de la misma y la necesidad de realizar procesos de análisis cada vez más complejos hace muy difícil la labor de generar nuevo conocimiento. Por tal razón, resulta de vital importancia para el investigador manejar las herramientas computacionales básicas que le permitan asegurar el éxito de sus proyectos de investigación.

OBJETIVO GENERAL

Introducir a los estudiantes universitarios, profesores de todos los niveles e investigadores del área de las ciencias biológicas en el manejo de las herramientas básicas de la bioinformática que provee el Centro Nacional para la Información en Biotecnología (NCBI) de los Estados Unidos de Norteamérica (<http://www.ncbi.nlm.nih.gov/>).

OBJETIVOS ESPECÍFICOS

Conocer las diversas aplicaciones diseñadas por NCBI para la búsqueda y manejo de información importante para la investigación en ciencias biológicas.

Aprender a manejar las bases de datos GenBank, MMDB y bases de datos derivadas de NCBI.

Aprender a manejar los programas de búsqueda de similitudes y otros recursos informáticos.

CONTENIDO DEL CURSO

Sesión 1(Octubre 11): Presentación

¿Que es la bioinformática?

¿Cómo y cuándo aparece?

Principales bases de datos en bioinformática ([NCBI](#), [EBI](#), [SIB](#), [EMBL](#))

La necesidad de la Bioinformática en Colombia (redes)

UNIX: El sistema operativo por excelencia en bioinformática.

Base de datos GenBank: descripción y alcances

Bases de datos derivadas de NCBI: RefSeqs

Búsqueda en bases de datos usando Entrez

- a. Información relacionada
- b. Búsqueda en Entrez

Sesión 2 (Octubre 12): Conceptos básicos de Biología Molecular

Estructura y composición del ADN y del ARN

Estructura y composición de proteínas

Dogma central de la Biología Molecular

Código genético

Estructura de los genes (promotores, amplificadores, represores)

Marcos de lectura abiertos

Mapas físicos y genéticos

Islas CpG

Bases de datos de estructuras de NCBI

- a. Base de datos de modelaje molecular (MMDB)
- b. Alineamientos estructurales
- c. Observación de estructuras y alineamientos estructurales con Cn3D

Sesión 3 (Octubre 13): Importancia de la comparación de secuencias genicas y protéicas

¿Porqué y cómo comparar secuencias?

Búsquedas de similitudes usando BLAST de NCBI

- a. Estadística de alineamientos locales
- b. Sistemas de conteo
- c. Uso de los servicios de BLAST en la web
- d. PSI-BLAST
- e. RPS-BLAST
- f. Páginas especializadas de BLAST

Sesión 4 (Octubre 14): Aplicación práctica de la bioinformática

Recursos para genómica en NCBI

- a. Genomas microbianos completos en Entrez
- b. Recursos para genomas de organismos superiores
 - RefSeq y genes
 - Unigene
 - Datos de variación (SNPs)
 - Los genomas del humano, ratón y rata
 - Visor de Mapas
 - Otros genomas

Trabajo complementario

Alineamiento múltiple

El problema del número de secuencias en el alineamiento múltiple

Encontrando zonas conservadas

Utilidad en el diseño de iniciadores o cebadores

Programas en INTERNET (uso en línea)

Programas para descarga y uso local

Construcción de dendrogramas

Métodos de distancia

Tipos de distancia: Modelos de sustitución (Ácidos nucleicos y proteínas)

Coordinación General:

Mauricio Pulido Jiménez. Coordinador CEBM Gimnasio Campestre

Javier Hernández Fernández. Asesor CEBM Gimnasio Campestre

Jaime Bernal Villegas, M.D., Ph.D. Rector Gimnasio Campestre

Conferencista:

Leonardo Mariño Ramírez, Ph.D.

Staff Scientist

National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

Bethesda, MD – USA

<http://www.ncbi.nlm.nih.gov/CBBresearch/Marino/>

Monitor:

Javier Hernández-Fernández, M.Sc. Asesor CEBM Gimnasio Campestre

Dirigido a:

El curso está abierto a estudiantes universitarios, profesores de todos los niveles e investigadores de cualquier área de las ciencias biológicas. Es recomendable conocimientos elementales de biología molecular, conocimientos básicos de manejo de computadores y conocimientos de inglés.

Fecha de realización:

Octubre 11 – 14 de 2005 (2-6 p.m)

Lugar:

Laboratorio de Informática, Gimnasio Campestre.

Calle 165 No. 19-50. Bogotá.

Cupo máximo:

30 personas

Valor de la inversión:

\$250.000

Informes e inscripciones:

Tel: 526-1700 / 526-1727 / 526-1737 ext. 269

Fax: 526-1710

E-mail: centrobiomol@campestre.edu.co

Curso Internacional

Manejo de Herramientas Básicas en Bioinformática

Conferencista

Leonardo Mariño Ramírez, Ph.D.
Staff Scientist
National Center for Biotechnology Information
National Library of Medicine
Bethesda, MD – USA

Octubre 2 – 3 de 2006
Universidad de La Sabana.
Campus Universitario. Puente del Común.
Chía. Autopista Norte Km 21.

PRESENTACIÓN

A pesar de ser la rama de las ciencias biológicas más recientemente surgida, la bioinformática es el área del conocimiento que ha hecho los aportes más valiosos a muchos de los grandes logros científicos de nuestro tiempo. Gracias a la aplicación de la informática a las ciencias biológicas algunas labores como la creación de bancos de información, el procesamiento y análisis de resultados experimentales y la comparación de grandes cantidades de datos son cada vez más rápidas y confiables. En la actualidad buena parte del avance de los proyectos de investigación en biología molecular, bioquímica, genética y biotecnología depende del uso de las herramientas que provee la bioinformática.

El inmenso caudal de información científica que se genera día a día, la profundidad de la misma y la necesidad de realizar procesos de análisis cada vez más complejos hace muy difícil la labor de generar nuevo conocimiento. Por lo tanto, resulta de vital importancia para el investigador manejar las herramientas computacionales básicas que le permitan asegurar el éxito de sus proyectos de investigación. Debido a esto, el presente curso busca introducir a los estudiantes universitarios, profesores e investigadores del área de las ciencias biológicas en el manejo de las herramientas básicas en Bioinformática que provee el Centro Nacional para la Información en Biotecnología (NCBI) de los Estados Unidos de Norteamérica.

Contenido del Curso

Día 1. Octubre 2 de 2006	
Conceptos básicos de Biología Molecular Estructura y composición del ADN, ARN y proteínas. Código genético. Estructura de los genes. Marcos de lectura abiertos. Mapas físicos y genéticos.	Sesión 1 8 a.m. - 10 a.m.
Refrigerio	10:00 a.m. - 10:30 a.m.
Introducción a la Bioinformática ¿Que es la bioinformática? ¿Cómo y cuándo aparece? Principales bases de datos en bioinformática: - NCBI, EBI, SIB, EMBL. Base de datos GenBank: descripción y alcances. Bases de datos derivadas de NCBI: RefSeqs. Búsqueda en bases de datos usando Entrez. - Información relacionada. - Búsqueda en Entrez.	Sesión 2 10:30 a.m. - 12:30 a.m.
Almuerzo	12:30 - 1:30
Bases de datos de estructuras del NCBI - Base de datos de modelamiento molecular (MMDB). - Alineamientos estructurales. - Observación de estructuras y alineamientos estructurales con Cn3D.	Sesión 3 1:30 p.m. - 3:30 p.m.
Refrigerio	3:30 p.m. - 4:00 p.m.
Comparación de secuencias. Parte I ¿Porqué y cómo comparar secuencias de genes y proteínas? Búsquedas de similitudes usando BLAST del NCBI. - Estadística de alineamientos locales. - Sistemas de conteo. - Uso de los servicios de BLAST en la web.	Sesión 4 4:00 p.m. - 6:00 p.m.

Contenido del Curso

Día 2. Octubre 3 de 2006	
Comparación de secuencias. Parte II - PSI-BLAST. - RPS-BLAST. - Páginas especializadas de BLAST. Alineamiento múltiple - El problema del número de secuencias. - Encontrando zonas conservadas.	Sesión 5 8 a.m. - 10 a.m.
Refrigerio	10:00 a.m. - 10:30 a.m.
Recursos para genómica en el NCBI Genomas microbianos completos en Entrez. Recursos para genomas de organismos superiores. - RefSeq y genes. - Unigene. - Datos de variación (SNPs). - Los genomas del humano, ratón y rata. - Visor de mapas.	Sesión 6 10:30 a.m. - 12:30 a.m.
Almuerzo	12:30 - 1:30
Diseño de iniciadores Conceptos básicos. Programas en línea (INTERNET). Programas para descarga y uso local.	Sesión 7 1:30 p.m. - 3:30 p.m.
Refrigerio	3:30 p.m. - 4:00 p.m.
Construcción de dendrogramas Conceptos básicos. Métodos de distancia. Tipos de distancia: modelos de sustitución (ácidos nucleicos y proteínas). Análisis de datos: RFLP, AFLP, RAPD, PFGE. Programas empleados.	Sesión 8 4:00 p.m. - 6:00 p.m.

Coordinación General

Yenny Milena Gómez P. Coordinadora de Investigación.
Facultad de Medicina, Universidad de La Sabana.
Javier Hernández-Fernández, Docente investigador.
Zootecnia, Universidad de La Salle

Conferencista

Leonardo Mariño Ramírez, Ph.D. NCBI. USA.
<http://www.ncbi.nlm.nih.gov/CBBresearch/Marino/>
Correo electrónico: marino@ncbi.nlm.nih.gov

Monitor

Elkin Hernández Porras, Médico Investigador.
Facultad de Medicina. Universidad de La Sabana.

Dirigido a

Estudiantes universitarios, profesores e investigadores de las diferentes áreas de las ciencias biológicas interesados en aprender el manejo de las herramientas básicas en Bioinformática.

Cupo limitado

Máximo 25 personas.

Valor de la inversión

\$250.000. **Descuento para estudiantes.**

El precio del curso incluye refrigerios y almuerzos.

Se entregará un certificado de asistencia al finalizar el curso.

Las inscripciones finalizan el 15 de septiembre de 2006.

Procedimiento de inscripción y pago

- 1-** Envíe un mensaje al correo electrónico investigacion.medicina@unisabana.edu.co con los siguientes datos: Nombres y apellidos completos (así aparecerán en el certificado de asistencia). Dirección y teléfono para contactar. Institución a la cual se encuentra vinculado.
- 2-** Inmediatamente recibirá un correo electrónico confirmando la reserva del cupo y los datos necesarios para el pago del curso mediante consignación bancaria.
- 3-** Envíe por fax o correo electrónico copia del recibo de la consignación.

4- En 1 a 3 días recibirá otro correo confirmando la inscripción al curso y los detalles concretos del lugar donde se realizará. Las inscripciones cuyo pago no sea realizado antes de iniciar el curso podrán ser anuladas.

Informes e inscripciones

Teléfono: 8615555 Ext. 2654 – 2605.

Fax: 8615555 Ext. 2626 – 2612.

Correo electrónico: investigacion.medicina@unisabana.edu.co
elkineh@yahoo.com

Introducción

Apreciando el creciente desarrollo de la Bioinformática y su amplia aplicabilidad en las investigaciones que involucran el manejo de datos biológicos, resulta indispensable conocer cada una de las herramientas computacionales disponibles y los alcances que estas puedan llegar a tener, comprender como utilizar los sistemas de cómputo facilitaría el análisis de las secuencias de ADN, de ARN, así como el análisis de la secuencia y estructura de las proteínas.

El curso brindará las pautas para que los investigadores de diferentes áreas de las ciencias biológicas, pueda localizar y emplear los recursos bioinformáticos, destinados para este fin.

Programa

Día 1. Junio 12

Introducción a la Bioinformática

¿Qué es Bioinformática?

Principales bases de datos en bioinformática: NCBI, EBI, SIB, EMBL

Manejo de ENTREZ, aplicaciones y búsqueda

Fundamentos de UNIX/Linux

Día 2. Junio

Alineamiento de secuencias

Búsqueda y alineamiento de secuencias

Alineamientos basados en estructuras 3D

Métodos de alineamiento de secuencias

BLAST

Almuerzo

Análisis de secuencias

Conceptos básicos

Diseño de primers

Herramientas para el diseño de PCRs

Programas para descarga y uso local

Día 3. Junio 14

Recursos del NCBI para genómica

Identificación de genes con herramientas del NCBI

Asignación de funciones para genes

Datos de variación (SNPs)

Análisis de información sobre SNPs, haplotipos y estudios de asociación

Almuerzo

Ácidos nucleicos y proteínas

Ácidos nucleicos - conceptos básicos

Micro-arreglos de ADN, qué son y cómo trabajan?

Aplicaciones

Proteínas - conceptos básicos

De la secuencia a la estructura

Predicción y caracterización de la función

Día 4. Junio 15

Construcción de dendrogramas (primera parte)

Conceptos básicos

Arboles filogenéticos ¿Qué son y cómo se interpretan?

Métodos de distancia

Almuerzo

Construcción de dendrogramas (segunda parte)

Modelos de sustitución (ácidos nucleicos y proteínas)

Análisis de datos: RFLP, AFLP, RAPD, PFGE

Programas de filogenia

Día 5. Junio 16

Consideraciones finales

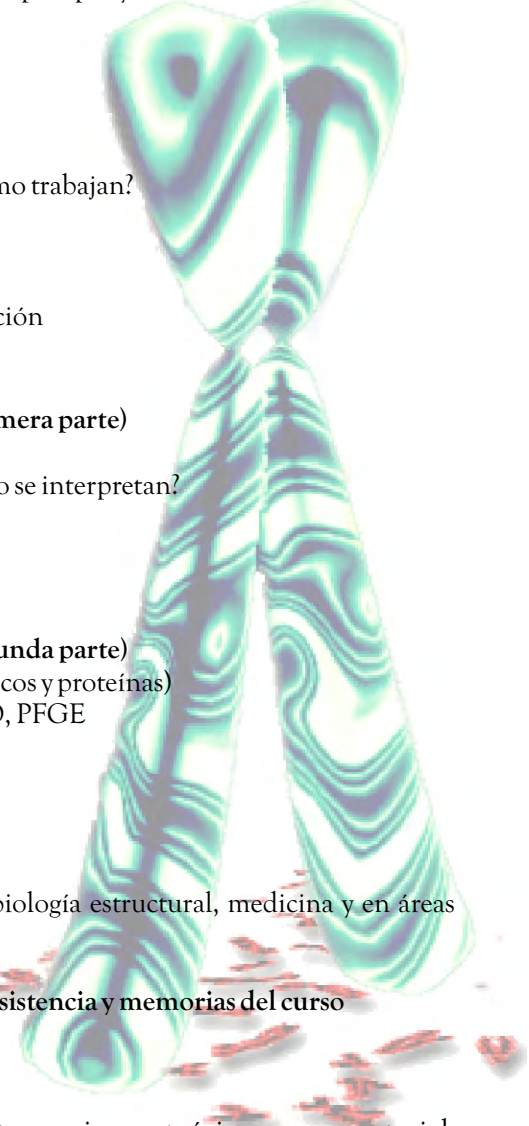
Genómica y Bioinformática

Aplicación de la bioinformática en biología estructural, medicina y en áreas afines

Clausura - entrega de certificado de asistencia y memorias del curso

Metodología

El curso se desarrollará en cuatro sesiones teóricas, con material audiovisual y de forma magistral e igual número de sesiones prácticas. Los alumnos realizarán tareas específicas complementarias a la instrucción teórica previa, valiéndose de los recursos bioinformáticos disponibles en Internet. El curso tendrá una intensidad de 24 horas.



Dirigido a

Estudiantes de pregrado y posgrado, docentes, investigadores y demás profesionales de las diferentes áreas de las ciencias biológicas, las matemáticas y la estadística involucrados en el manejo de datos biológicos e interesados en el manejo de las principales herramientas bioinformáticas.

Inversión

\$300.000. El valor del curso incluye el derecho a refrigerios, memorias y certificado de asistencia. CUPO LIMITADO.

Patrocina



Organiza

Grupo de Investigaciones Biomédicas y Biología Molecular
Facultad de Ciencias de la Salud - UNISINU

Coordinadores:

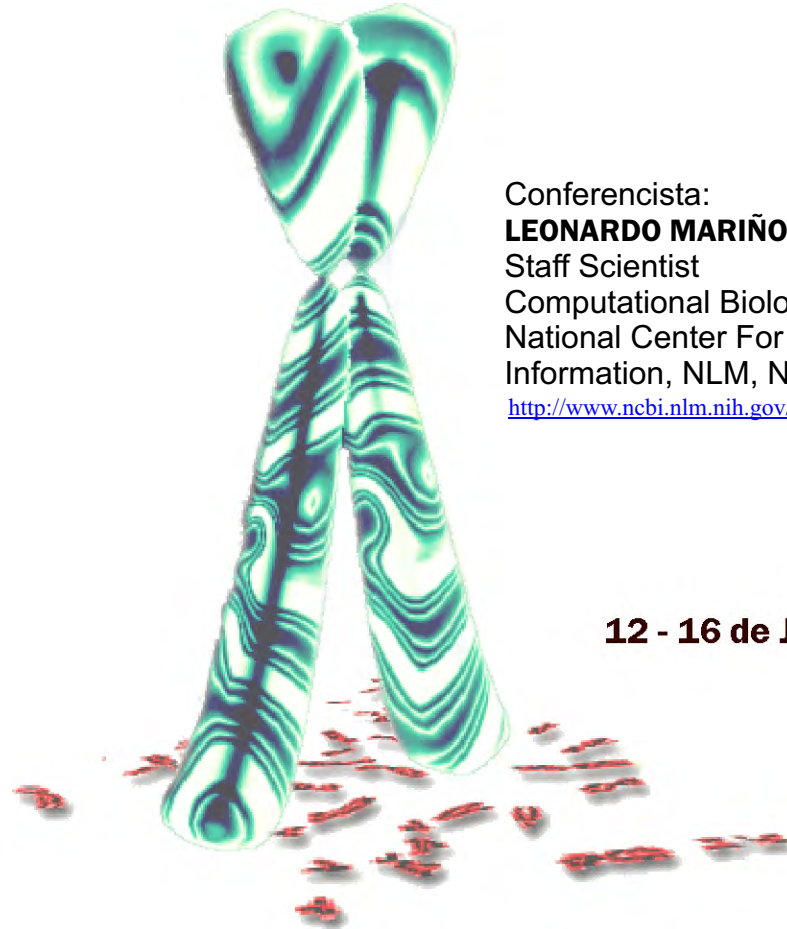
Milton Quintana Sosa
Vanessa Otero Jiménez
Manolo Jaramillo García

Contacto: Telefax (4) 784 19 61

Email: cursobioinformatica2007@unisinu.edu.co

CURSO DE BIOINFORMATICA

“Fundamentos para el manejo y uso de datos Biológicos”



Conferencista:

LEONARDO MARIÑO-RAMIREZ Ph.D.

Staff Scientist

Computational Biology Branch
National Center For Biotechnology
Information, NLM, NIH.

<http://www.ncbi.nlm.nih.gov/CBBresearch/Marino/>

12 - 16 de Junio de 2007



UNIVERSIDAD DEL SINU

Elías Bechara Zainúm



GRUPO DE INVESTIGACION
BIOMEDICA Y BIOLOGIA MOLECULAR

I Curso Internacional de Bioinformática Básica.

Porque primero se debe entender lo elemental

Octubre 24 y 25 de 2011

Si se siente identificado con estas frases, este curso es para usted:

- No se que hacer con toda la información que arroja cualquier búsqueda en GenBank.
- Necesito saber las características bioquímicas más básicas de una secuencia proteica y no se por donde empezar.
- De todas las secuencias que existen de una misma proteína, hay alguna que sea de referencia?
- Me es confuso entender cualquier resultado de un Blast. Es mas, no se que tipo de BLAST usar.
- No se cual es la diferencia entre alineamiento local y alineamiento global.
- No se como seleccionar el mejor resultado de una análisis de biología computacional.
- No se como interpretar un SNP y sus haplotipos en el contexto de una enfermedad humana.
- No se por donde empezar a estudiar los epítopes en un candidato a vacuna.
- No se como modelar la estructura 2D y 3D de una secuencia de proteínas.
- Como me puedo dar cuenta de que mutaciones se asocian a que enfermedades?
- Como debo interpretar la información del OMIM y del dbSNP?
- Me es difícil usar Pubmed para búsqueda de documentos y referencias.
- No se como acceder a los artículos científicos de manera gratuita.

Dirigido a:

* Docentes universitarios y estudiantes de pregrado, maestría o doctorado que necesiten ayuda de carácter fundamental en biología computacional y bioinformática para sus proyectos de investigación.

* Estudiantes de pregrado y postgrado en áreas de la salud humana interesados en saber como usar adecuadamente los recursos informáticos más importantes que existen actualmente en este campo.

Temas:

- Genomas, secuencias y mapas.
- Proteínas dominios y estructuras.
- NCBI Blast y alineamientos múltiples.
- Herramientas de predicción de epítopes
- Variación humana y enfermedad.
- Al final se contará con un espacio para asesoría específica a problemas de investigación

Organizado por:

•Grupo de Investigación Instituto de Investigaciones Biomédicas.



Patrocinado por:
Universidad Libre – Seccional Cali.



Docentes:

- King Jordan, PhD.
- Associate Professor, School of Biology.
- Georgia Institute of Technology. Atlanta, Georgia (USA).
- And co-founder of the PanAmerican Bioinformatics Institute

- Leonardo Mariño, PhD.
- Staff Scientist, Computational Biology Branch.
- NCBI, NLM, NIH - Bethesda, Maryland (USA).
- And co-founder of the PanAmerican Bioinformatics Institute

- Augusto Valderrama-Aguirre, MSc. PhDc.
- Biología Molecular e Inmunología.
- Director Grupo Instituto Investigaciones Biomédicas.
- Docente asociado, Facultad Salud, Universidad Libre-Cali (Colombia)

INFORMES:

iib.bioinformatica2011@gmail.com

Inversión:

- Docentes: \$ 400.000
- Profesionales: \$ 400.000
- Estudiantes de posgrado: \$ 300.000
- Estudiantes de pregrado : \$ 250.000
- Cupos limitados

•Procedimiento Inscripción:

- Diligenciar el formulario disponible en la página web de la universidad (www.unilibrecali.edu.co/Bioinformatica).
- Consignar en la cuenta de ahorros 30410084886 de Bancolombia (con nombre y cédula del inscrito) y enviar la consignación por correo electrónico a avalderr@hotmail.com

V Simposio de Biología Molecular

**Diagnóstico de Acompañamiento
La Revolución Genómica Aplicada a la
Medicina Moderna**

Lugar: Promédico

Avenida 6 AN. No. 22N-54

sábado 22 de mayo de 2010

hora: 8:00 a.m. a 6:30 p.m.

Invitados Especiales:

***Dr. ÁLVARO JOSÉ GUERRERO, MD (Colombia)**
Especialista en Oncología Clínica.

***Dra. MARÍA LUISA MAESTRO, PhD (España)**
Especialista en Biomarcadores Tumorales

***Dr. LEONARDO MARIÑO RAMÍREZ, PhD (USA)**
Biólogo Computacional.



**Universidad Libre
Seccional Cali
Facultad de Ciencias de la Salud**

Informes:

Celular: 316 582 8097

Correo: katicol7@hotmail.com

Patrocinadores:

Organiza:
3° Semestre de Medicina y Cirugía, apoyado por
el grupo Instituto de Investigaciones Biomedicas



V Simposio Biología Molecular

DIAGNÓSTICOS DE ACOMPAÑAMIENTO LA REVOLUCIÓN GENÓMICA APLICADA A LA MEDICINA MODERNA.

Universidad Libre-Seccional Cali
Instituto de Investigaciones Biomédicas
III Semestre Medicina y Cirugía

PROGRAMA

SESION 1. INTRODUCCIÓN A LOS DIAGNÓSTICOS DE ACOMPAÑAMIENTO Y MEDICINA PERSONALIZADA.

Hora	Tema	Conferencista(s)
8:00 – 9:00 am (60 min)	Videoconferencia internacional: Marcadores Moleculares en Tumores Sólidos.	<i>Dra María Luisa Maestro de las Casas, PhD</i> (España). Laboratorio Patología Molecular. Hospital Clínico San Carlos – Madrid.
9:00 – 9:20 am (20 min)	– Bienvenida. – Introducción: Medicina personalizada, diagnósticos de acompañamiento e investigación transaccional.	<i>Augusto Valderrama Aguirre MSc. PhD(C)</i> Profesor Asociado, Universidad Libre-Cali Director Grupo IIB
9:20 – 10:00 am (40 min)	Conferencia Magistral I: Genomas virales vs genoma humano.	<i>Felipe García Vallejo, PhD</i> Profesor Titular, Universidad del Valle Director Laboratorio de Biología Molecular y Patogénesis.
9:40 – 10:00 am (20 min)	<i>Coffee Break</i> (30 min)	

SESION 2. IDENTIFICACIÓN MOLECULAR DE SUJETOS CON EL GENOTIPO ADECUADO PARA EL TRATAMIENTO.

Hora	Temas	Conferencista(s)
10:00 – 10:40 am (40 min)	Conferencia Magistral II: Tecnologías de última generación para la determinación de marcadores genómicos en medicina personalizada.	<i>Dr Leonardo Mariño, PhD</i> (USA). Computational Biology Branch. National Center for Biotechnology Information (NCBI/NLM/NIH).
10:40 – 11:05 am (25 min)	1. EGFR (Erlotinib y Gefitinib). 2. KIT (Imatinib).	<i>Grupo 1 – Estudiantes 3er semestre Medicina y Cirugía, Universidad Libre – Cali.</i>
11:05 – 11:30 am (25 min)	3. VEGFA (Bevacizumab). 4. ERBB2 (Trastuzumab).	<i>Grupo 2 – Estudiantes 3er semestre Medicina y Cirugía, Universidad Libre – Cali.</i>
11:30 – 11:55 am (25 min)	5. KRAS (Cetuximab). 6. NAT2 (Isoniazida). 7. VKORC1 (Warfarina)	<i>Grupo 3 – Estudiantes 3er semestre Medicina y Cirugía, Universidad Libre – Cali.</i>
11:55 am – 2:00 pm (2 horas)	<i>Espacio para almuerzo</i> (2 horas)	

SESION 3. IDENTIFICACIÓN MOLECULAR DE SUJETOS CUYA RESPUESTA AL TRATAMIENTO SEA POBRE O TÓXICA.

Hora	Temas	Conferencista(s)
2:00 – 2:35 pm (35 min)	Conferencia Magistral III: Aplicaciones de los Diagnósticos de Acompañamiento en Oncología Clínica y necesidades más imperantes en el suroccidente Colombiano.	Dr Álvaro Guerrero, MD (Colombia). Especialista en Hematología y Oncología. Hemato-oncólogos de Occidente.
2:35 – 3:00 pm (25 min)	1. CYP2C19 (Omeprazol). 2. CYP2C9 (Warfarina). 3. CYP2D6 (Fluoxetina).	Grupo 4 – Estudiantes 3er semestre Medicina y Cirugía, Universidad Libre – Cali.
3:00 – 3:25 pm (25 min)	4. HLA-B (Abacavir, Carbamazepina). 5. DPYD (Capecitabina, 5-Fluorouracilo).	Grupo 5 – Estudiantes 3er semestre Medicina y Cirugía, Universidad Libre – Cali.
3:25 – 3:50 pm (25 min)	6. G6PD (Cloroquina, Dapsona). 7. TPMT (Mercaptopurina). 8. UGT1A1 (Irinotecan).	Grupo 6 – Estudiantes 3er semestre Medicina y Cirugía, Universidad Libre – Cali.
3:50 – 4:15 pm (25 min)	<i>Coffee Break</i> (30 min)	

SESION 4. TECNOLOGÍAS DE APROBADAS POR LA FDA PARA DIAGNÓSTICO DE ACOMPAÑAMIENTO.

Hora	Temas	Conferencista(s)
4:15 – 4:40 pm (25 min)	AmpliChip Cytochrome P450 (Roche).	Grupo 7 – Estudiantes 3er semestre Medicina y Cirugía, Universidad Libre – Cali.
4:40 – 5:05 pm (25 min)	KRAS mutation detection kit (Qiagen)	Grupo 8 – Estudiantes 3er semestre Medicina y Cirugía, Universidad Libre – Cali.
5:05 – 5:30 pm (25 min)	1. PathVysion HER-2 DNA Probe Kit (Abbott) 2. TheraGuide (Myriad)	Grupo 9 – Estudiantes 3er semestre Medicina y Cirugía, Universidad Libre – Cali.
5:30 – 6:00 pm (25 min)	<i>Brindis de cierre</i> (30 min)	



UNIVERSIDAD NACIONAL DE COLOMBIA

VICERRECTORÍA DE INVESTIGACIÓN
DIRECCIÓN DE INVESTIGACIÓN SEDE BOGOTÁ

DIB-719-2011

Bogotá, 9th September 2011

Professor
LEONARDO MARIÑO RAMIREZ
NCBI, NLM, NIH
Computational Biology Branch
Bethesda, MD
United States of America

Dear Professor Mariño:

The Academic and Research Directors at the National University of Colombia are pleased to announce the upcoming **International school**, which will be offered this summer and a Special Course on November. The course will be part of a series of lectures given by international experts on new technologies in health and Eco-cities.

The International School will intend to cover several topics about these fields. Specifically, the topic of new technologies in health will have a section focused on **Genomics in Infectious Diseases, a public tool** (this course will be developed, from Nov 28th to Dic. 2nd, 2011). Given your professional and academic achievement and your contributions to the field, it's a pleasure for the International School to invite you to participate as a guest speaker and to host you in our city.

According to the program of the International School, your proposed lectures will be on:

- A. Genomics, genome sequencing and annotation. Tools overview: Next-generation sequencing and Genome assembly.
- B. Practical Session 1 B: The genome sequence of *Mycobacterium colombiense* CECT 3035 type strain - Genome assembly and annotation Demo

We will be grateful for your visit at least during the following days. The National University of Colombia will cover flight tickets (in economic class) and the hotel, while you are in this event. No US Federal Government funds will be used to pay for any portion of the travel and no honorarium will be given to the traveler.

Sincerely yours,

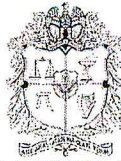
LUIS FERNANDO NIÑO VÁSQUEZ
Research Director

JUAN MANUEL TEJEIRO
Academic Director

Carolina G

ciencia, tecnología e innovación para el país

Carrera 45 No. 26-85, **EDIFICIO URIEL GUTIÉRREZ**, 2º piso Oficina 206
Teléfono: (57-1) 316 5099 Conmutador: (57-1) 316 5000 Ext. 18115 – 18151 - 18163 Fax: Ext. 18172
Correo electrónico: dirinvesti_bog@unal.edu.co / Bogotá, Colombia, Sur América



UNIVERSIDAD NACIONAL DE COLOMBIA
CONSEJO DE FACULTAD - MEDICINA

RESOLUCIÓN NÚMERO: 768 DE 28 de Mayo de 2010

"Por la cual se autoriza la vinculación de un Docente Ocasional Ad-honorem"

EL DECANO

En uso de sus facultades legales y estatutarias

CONSIDERANDO:

Que el Acuerdo del Consejo Superior Universitario No. 016 de 2005, estableció el reglamento del personal académico no vinculado a la carrera Profesoral Universitaria y los requisitos para su contratación en la Universidad Nacional de Colombia.

Que el numeral 6 del artículo 24 del citado Acuerdo, establece que se podrán contratar docentes Ocasionales temporalmente, con el fin de desarrollar actividades docentes en la Universidad.

Que el Parágrafo No. 1, del Numeral 6, del mismo Artículo, establece que los Profesores Visitantes, Los Profesores Especiales y los Profesores Ocasionales podrán ser vinculados Ad-honorem.

Que la Facultad De Medicina, seleccionó de conformidad con el procedimiento establecido en las disposiciones vigentes, al profesor **LEONARDO MARIÑO RAMÍREZ**, identificado con cédula de ciudadanía No. 79.520.882 expedida en Bogotá, en la modalidad de Docente Ocasional Ad-honorem con una dedicación de 12 horas semanales, para dictar la(s) asignatura(s) Asesorar la tesis de la estudiante Mónica Gonzales del doctorado en Ciencias Biomédicas, en el tema "secuencia de genoma de Mycobacterium Colombiense" - Organización y planeación de Curso de informática, durante el período comprendido entre el 1 de junio de 2010 y el 17 de diciembre de 2010, en el Departamento de Microbiología.

RESUELVE:

ARTÍCULO 1°.- Autorizar la vinculación en la modalidad de Docente Ocasional Ad-honorem de el profesor **LEONARDO MARIÑO RAMÍREZ** identificado con cédula de ciudadanía No. 79.520.882 expedida en Bogotá, con una dedicación de 12 horas semanales, para dictar la(s) asignatura(s) Asesorar la tesis de la estudiante Mónica Gonzales del doctorado en Ciencias Biomédicas, en el tema "secuencia de genoma de Mycobacterium Colombiense" - Organización y planeación de Curso de informática, durante el período comprendido entre el 28 de Mayo 2010 y el 17 de diciembre de 2010, en el Departamento de Microbiología.

ARTÍCULO 2°.- La presente vinculación no genera derechos de carrera.

ARTÍCULO 3°.- Envíese copia de la presente Resolución a la División de Personal Académico.

ARTÍCULO 4°.- Por Secretaría de Facultad, notificar del contenido de la presente Resolución al profesor **LEONARDO MARIÑO RAMÍREZ**.

ARTÍCULO 5°.- La presente resolución rige a partir de la fecha de expedición.

COMUNÍQUESE Y CÚMPLASE

Dada en Bogotá D.C., el 28 de Mayo de 2010

CARLOS JULIO PACHECO CONSUEGRA
DECANO

J. Pincedillo

4

Favre
09/06/10



UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE MEDICINA

RESOLUCIÓN NÚMERO: 00278 DE 23 de febrero de 2011

"Por la cual se autoriza la vinculación de un Docente Ocasional Ad-honorem"

EL DECANO DE LA FACULTAD DE MEDICINA

En uso de sus facultades legales y estatutarias

CONSIDERANDO:

Que el Acuerdo del Consejo Superior Universitario No. 016 de 2005, estableció el reglamento del personal académico no vinculado a la carrera Profesorial Universitaria y los requisitos para su contratación en la Universidad Nacional de Colombia.

Que el numeral 6 del artículo 24 del citado Acuerdo, establece que se podrán contratar docentes Ocasionales temporalmente, con el fin de desarrollar actividades docentes en la Universidad.

Que el Parágrafo No. 1, del Numeral 6, del mismo Artículo, establece que los Profesores Visitantes, Los Profesores Especiales y los Profesores Ocasionales podrán ser vinculados Ad-honorem.

Que la Facultad de Medicina, seleccionó de conformidad con el procedimiento establecido en las disposiciones vigentes, al profesor **LEONARDO MARIÑO RAMÍREZ**, identificado con cédula de ciudadanía No. 79.520.882 expedida en Bogotá, en la modalidad de Docente Ocasional Ad-honorem con una dedicación de 12 horas semanales, para dictar la(s) asignatura(s) ASESORIA DE TESIS ESTUDIANTE DOCTORADO, durante el período comprendido entre el 24 de febrero de 2011 y el 17 de diciembre de 2011, en el Departamento de Microbiología.

RESUELVE:

ARTÍCULO 1°.- Autorizar la vinculación en la modalidad de Docente Ocasional Ad-honorem de el profesor **LEONARDO MARIÑO RAMÍREZ** identificado con cédula de ciudadanía No. 79.520.882 expedida en Bogotá, con una dedicación de 12 horas semanales, para dictar la(s) asignatura(s) ASESORIA DE TESIS ESTUDIANTE DOCTORADO, durante el período comprendido entre el 24 de febrero de 2011 y el 17 de diciembre de 2011, en el Departamento de Microbiología.


ARTÍCULO 2°.- La presente vinculación no genera derechos de carrera.

ARTÍCULO 3°.- Envíese copia de la presente Resolución a la División de Personal Académico.


ARTÍCULO 4°.- Por Secretaría de Facultad, notificar del contenido de la presente Resolución al profesor **LEONARDO MARIÑO RAMÍREZ**.

ARTÍCULO 5°.- La presente resolución rige a partir de la fecha de expedición.

NOTIFIQUESE, COMUNÍQUESE Y CÚMPLASE
Dada en BOGOTÁ D.C., el 23 de febrero de 2011


CARLOS ALBERTO AGUDELO CALDERON
DECANO

Jeannette P.


Jeannette P.
28/02/11



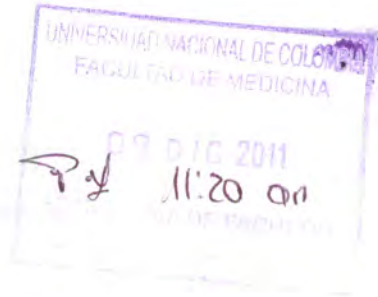
6758-000

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ
FACULTAD DE MEDICINA
DOCTORADO EN CIENCIAS BIOMÉDICAS - DCBM

DCBM 120-11

Bogotá, 5 de diciembre de 2011

Señores
COMITÉ DE DIRECTORES
Facultad de Medicina
Universidad Nacional de Colombia
Ciudad



Respetados Señores:

Atentamente solicito a ustedes el nombramiento del Doctor **LEONARDO MARIÑO RAMÍREZ**, identificado con la Cédula de Ciudadanía No 79.520.882 de Bogotá, en la modalidad de Docente Ocasional Ad-honorem con una dedicación de 12 horas semanales, para asesorar la Tesis de la estudiante del Programa de Doctorado en Ciencias Biomédicas Mónica Natalia González Pérez con el tema: **"RESPUESTA INMUNE Y VIRULENCIA DE mycobacterium colombiense: UN POSIBLE PATÓGENO OPORTUNISTA EMERGENTE EN COLOMBIA"**; Organización y planeación de Curso de informática, durante el periodo comprendido entre el 16 de enero y el 17 de diciembre de 2012, en el Programa de Doctorado en Ciencias Biomédicas.

Agradeciendo su amable colaboración,

Atentamente

CARLOS ALBERTO PARRA LÓPEZ
Coordinador Doctorado en Ciencias Biomédicas

TC

ciencia, tecnología e innovación para el país

Calle 44 No. 45-67, **UNIDAD CAMILO TORRES**, Módulo 2 Bloque C Oficina 901
Conmutador: (57-1) 316 50 00 Ext. 10921 – 15039 Telefax: ext. 10922
Correo electrónico: docbiomed_fmbog@unal.edu.co
Bogotá, Colombia, Sur América



INTERNATIONAL CENTRE FOR GENETIC ENGINEERING AND BIOTECHNOLOGY

Theoretical and Practical Course "Bioinformatics: Computer Methods in Molecular and Systems Biology"

ICGEB and AREA Informatics Lab, AREA Science Park
Trieste, Italy, 25-30 June 2012

PRELIMINARY PROGRAMME

MONDAY, 25 June

08:00	Pick-up from Hotel Roma	
08:30	Registration	<i>ICGEB Foyer, 'W' Building</i>
09:15	Welcome Address	<i>ICGEB Seminar Room, 'W' Building</i> Sándor Pongor , ICGEB, Trieste, Italy
09:30	Bioinformatics and knowledge representation: theoretical intro to the course	Sándor Pongor
10:45	Break	
11:15	The ICGEB computer system, course infrastructure	<i>Informatics Laboratory, 'E3' Building</i> Dario Palmisano , ICGEB, Trieste, Italy Sándor Pongor
12:30	Sequence analysis, database searching	David Judge , University of Cambridge, UK
13:30	Lunch	<i>Cafeteria, Ground Floor, 'C' Building</i>
14:30	Sequence analysis, database searching (cont.)	David Judge
18:00	Get-Together Party	<i>ICGEB Foyer, 'W' Building</i>
19:00	Bus to Hotel Roma	

TUESDAY, 26 June

08:30	Pick-up from Hotel Roma	
09:00	Sequence analysis, database searching (cont.)	David Judge
11:00	Break	
11:30	Sequence analysis, database searching (cont.)	David Judge
13:30	Lunch	<i>Cafeteria, Ground Floor, 'C' Building</i>
14:30	Sequence analysis, database searching (cont.)	David Judge
18:00	Bus to Hotel Roma	

WEDNESDAY, 27 June

08:30	Pick-up from Hotel Roma	
09:00	Sequence analysis, database searching (cont.)	David Judge
11:00	Break	
11:30	Sequence analysis, database searching (cont.)	David Judge
13:30	Lunch	<i>Cafeteria, Ground Floor, 'C' Building</i>
14:30	KEGG: Kyoto Encyclopedia of Genes and Genomes	Minoru Kanehisa , Kyoto University and University of Tokyo, Japan
18:00	Bus to Hotel Roma	

THURSDAY, 28 June

08:30	Pick-up from Hotel Roma	
09:00	The European Bioinformatics Institute (EBI) services: searches, functional genomics and pathway databases	Gabriella Rustici , EBI, Hinxton, UK
11:00	Break	
11:30	The EBI services (cont.)	Gabriella Rustici
13:30	Lunch	<i>Cafeteria, Ground Floor, 'C' Building</i>
14:30	The EBI services (cont.)	Gabriella Rustici
18:00	Bus to Hotel Roma	

FRIDAY, 29 June

08:30	Pick-up from Hotel Roma	
09:00	The National Center for Biotechnology Information (NCBI) services, nucleic acid databases	Leonardo Marino Ramirez , NCBI , NIH, Bethesda, MD, USA
11:00	Break	
11:30	NCBI services (cont.)	Leonardo Marino Ramirez
13:30	Lunch	<i>Cafeteria, Ground Floor, 'C' Building</i>
14:30	Next-generation sequencing data analysis using open source software	Leonardo Marino Ramirez
18:00	Bus to Hotel Roma	

SATURDAY, 30 June

08:30	Pick-up from Hotel Roma	
09:00	Genome computing, phylogeny	Martin Bishop , Cambridge, UK, David Judge
10:30	Break	
11:00	Genome computing, phylogeny (cont.)	Martin Bishop , David Judge
13:00	Lunch	<i>ICGEB premises, 'W' Building</i>
14:00	Protein sequence databases, UniProt	Elisabeth Gasteiger , SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
18:00	Bus to Hotel Roma	

Boston University Bioinformatics Graduate Program

24 Cummington Street
Boston, Massachusetts 02215
T 617-358-0752 F 617-353-4814
www.bu.edu/bioinformatics



July 28, 2011

Leonardo Marino-Ramirez, Ph.D.

NCBI / NLM / NIH

Building 38A, Room 6S614M

8600 Rockville Pike

Bethesda, MD 20894

Dear Dr. Marino-Ramirez,

I am pleased to confirm your appointment as Adjunct Professor of the Bioinformatics Graduate Program at Boston University, effective September 1, 2011 through August 31, 2014. Thank you for submitting a request for appointment and the required documentation.

We would appreciate it if you keep the Program informed of major achievements in your scientific life and provide electronic copies of your papers for our files. Note that you will be receiving the Bioinformatics Program Newsletter on a regular basis and are more than welcome to submit items about your research and other activities – and news about the accomplishments of your students, both past and present.

Please accept my best wishes for successful participation in the Program. If you need any administrative support, please contact Caroline Lyman at clyman@bu.edu.

With best wishes,

A handwritten signature in black ink, appearing to read "Thomas D. Tullius".

Thomas D. Tullius

Director,

Boston University Bioinformatics Program

JOHNS HOPKINS UNIVERSITY
Zanvyl Krieger School of Arts and Sciences
Advanced Academic Programs
1717 Massachusetts Avenue NW, Suite 104
Washington, DC 20036

MEMORANDUM OF APPOINTMENT

Your teaching appointment in Advanced Academic Programs is indicated as follows:

Name: Dr. Leonardo Marino-Ramirez, Ph.D.
Address: 11129 Schuylkill Road
Rockville, MD 20852

DUPLICATE

Course(s) and Schedule:

Fall 2012 Term
410.666.71 Genomic Sequencing and Analysis
S 9:00 PM - 4:00 PM 9/8/2012 - 12/15/2012 Campus Location: MCC
Salary: \$3347, 6 semi-monthly payments beginning October 15th

Instructional appointments such as this are made exclusively on a per-course basis for a given term. Please note that this is a casual appointment for which there are no fringe benefits. The University reserves the right to cancel courses or sections or to adjust instructional salaries at its sole discretion. Should this latter action become necessary, you will be given the opportunity to decline to teach at a reduced salary. **Payments will begin on October 15th provided all paperwork is received by University Payroll by August 30th.**

In accordance with the regulations of the Advanced Academic Programs (AAP), the faculty member agrees to: provide a syllabus for the above-named course two weeks before the beginning of the semester to the site coordinator at your teaching site or posted to your Course Management System site for online courses; notify the site coordinator and the Program Director of any change in class time or place including rescheduling of any class cancelled for any reason. **Please read, sign, and return the enclosed Acknowledgement Regarding Textbook Selection form along with this signed appointment letter.** Design your course to include objective student evaluation and complete the mid-term student evaluation for any student in danger of receiving a B- or below at the mid-term point. All AAP students now complete their evaluations online, and receive the evaluation via email. Please encourage students to participate in online course evaluations, usually during the last two weeks of the semester. Once the results of the evaluations are summarized, faculty will be sent their individual course results.

Faculty are expected to meet all the scheduled class times, grade assignments, quizzes, and exams promptly with constructive and thoughtful feedback, and use their best professional efforts in developing and delivering their course. In the event that these conditions are not met, the university retains the right to take appropriate action up to and including removing the faculty member from the course.

We are pleased to welcome you as a lecturer. Your signature below confirms your willingness to accept this appointment. Please make a copy of this appointment letter for your records and return the original in the enclosed postage-paid envelope to **KSAS Human Resources, Wyman 6th Floor, Room S604, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218.**

Signature of Faculty

Kathleen M. Burke 8/16/2012
Kathleen M. Burke, Ph.D.
Associate Dean, Graduate Professional Programs

Proyectos de Investigación en Colombia

1. Secuencia del transcriptoma de la Uchuva *Physalis peruviana*
2. Secuencia del genoma de *Mycobacterium colombiense*

BioProject

BioProject (Geno

Search

[Limits](#) [Advanced](#)

[Help](#)

[Display Settings:](#)

[Send to:](#)

Name: Mycobacterium colombiense CECT 3035

Title: Mycobacterium colombiense CECT 3035 genome sequencing project

Accession: PRJNA67689 ID: 67689

Mycobacterium colombiense CECT 3035 is the type strain and is part of the Mycobacterium avium Complex. Since its description in 2006, it has been associated with opportunistic disease in immunosuppressed HIV-infected patients and causing lymphadenopathy in immunocompetent children.

Project data type: Genome sequencing

Attributes: Scope: Monoisolate; Material: Genome; Capture: Whole; Method type: Sequencing;

Lineage: Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Mycobacteriaceae; Mycobacterium; Mycobacterium avium complex (MAC); Mycobacterium colombiense; Mycobacterium colombiense CECT 3035

Project Data

Nucleotide: 25

Protein: 5230

WGS prefix: AFVW

Submission:

Registration date: 3-Jun-2011

Universidad Nacional de Colombia

- Emory GRA Genome Center
- National Center for Biotechnology Information
- PanAmerican Bioinformatics Institute
- Georgia Institute of Technology

Entrez Links

Nucleotide (25)

Protein (5230)

Recent activity

[Turn Off](#) [Clear](#)

Your browsing activity is empty.

Related information

Nucleotide

Project

Protein

Taxonomy

Umbrella Projects

Umbrella Project Title			Number of Subprojects
Mycobacterium colombiense (organism overview)			▼ 2
BioProject accession	Project type	Organism	Title
✓ PRJNA67689	Genome sequencing	Mycobacterium colombiense CECT 3035	Mycobacterium colombiense CECT 3035 genome sequencing project (Universidad Nacional de Colombia)
PRJNA71463	RefSeq Genome	Mycobacterium colombiense CECT 3035	Mycobacterium colombiense CECT 3035 genome sequencing project (NCBI)

ATPasas TIPO P DE *M. tuberculosis* COMO POSIBLES DIANAS TERAPÉUTICAS

En colaboracion con Carlos Yesid Soto Ospina, PhD

Durante la infección, el bacilo tuberculoso enfrenta condiciones adversas como una escasa disponibilidad de iones esenciales y de nutrientes necesarios para su supervivencia [1-5], así como un arsenal de sustancias tóxicas usadas por las células fagocíticas para eliminar patógenos intracelulares, entre ellas se pueden mencionar: altas concentraciones de H^+ , hidrolasas, péptidos antimicrobianos, especies reactivas de nitrógeno y oxígeno y altas concentraciones de cationes de metales pesados [6].

Probablemente, el éxito del bacilo tuberculoso durante la infección, en parte puede deberse a que responde de manera efectiva a este microambiente y se adapta a él. En ello, las ATPasas tipo P deben ser indispensables, ya que no solo transportan cationes metálicos, sino que generan los gradientes electroquímicos necesarios para el transporte de otro tipo de solutos a través de membranas celulares y también protegen a la célula de sustancias tóxicas presentes en los fagosomas [7, 8].

Las ATPasas tipo P son proteínas de membrana conservadas en estructura tridimensional, más que en secuencia de aminoácidos. Poseen de 6 a 10 TMS que orientan los extremos N- y C-terminales al lado citoplasmático de la membrana. El extremo N-terminal forma la gran cabeza citoplasmática caracterizada por formar 2 grandes loops en los que se ubican los principales sitios activos de la enzima. Por su parte, el extremo C-terminal está embebido casi en su totalidad en la membrana. Exhiben 5 dominios funcional y estructuralmente diferentes: tres citoplasmáticos (A, actuador; N, unión a nucleótido; y P, fosforilación) y dos embebidos en la membrana (T, transporte; y S, soporte específico de la clase) [7, 9-11]. Durante cada ciclo catalítico el dominio P es fosforilado en un residuo de aspartato, por acción del dominio N (actividad proteína-quinasa) y posteriormente defosforilado por el dominio A (actividad fosfatasa) [7, 9].

M. tuberculosis posee 12 ATPasas tipo P (CtpA, CtpB, CtpC, CtpD, CtpE, CtpF, CtpG, CtpH, CtpI, CtpJ, CtpV y KdpB), un gran número en comparación con micobacterias saprófitas como *M. smegmatis* que solo posee 6 ATPasas tipo P. 7 de esas doce ATPasas han sido postuladas como posibles transportadores de cationes de metales pesados [12], lo que sugiere una posible importancia de estas en la defensa celular de las micobacterias durante la infección, como se ha propuesto para otros procariotas y algunos eucariotas unicelulares [8]. Los metales pesados, que bajo ciertas condiciones son fundamentales para la vida, en concentraciones elevadas se convierten en agentes tóxicos que bloquean grupos funcionales, desplazan iones metálicos esenciales o modifican las conformaciones activas de moléculas biológicas [13]. Por ello, resulta fundamental que las bacterias sean capaces de establecer un equilibrio entre la entrada y la salida de cada metal, de tal forma que ellos se mantengan en la concentración a la que son nutrientes. Por lo tanto, los transportadores que permiten mantener dicha homeostasis deben ser fundamentales para la virulencia de estos patógenos intracelulares [14].

Acorde con la hipótesis anterior, se ha propuesto que CtpC, CtpG y CtpV, pueden hacer parte de un mecanismo de defensa usado por las micobacterias para sobrevivir por largos periodos de tiempo dentro de células fagocíticas humanas, ya que establecieron que se sobreexpresan durante el proceso de infección [6]. Actualmente se sabe que CtpV es un exportador de cobre necesario para la virulencia de *M. tuberculosis* [14], mientras que CtpC está involucrada en el eflujo de Zn^{2+} [6].

Recientemente se han desarrollado nuevos medicamento anti-TB, tales como diarilquinolinas y benzotiazinas cuyas dianas se ubican en la membrana plasmática micobacteriana [6] lo que muestra una ventaja, ya que antimicrobianos diseñados contra dianas en la membrana evitan los problemas relacionados con la permeabilidad de dicha barrera biológica. Uno de dichos antibióticos es un inhibidor de la F_1F_0 -ATP sintasa micobacteriana [15], otro tipo de ATPasa que al dejar de funcionar limita la cantidad de ATP disponible en la célula, es por ello que este medicamento puede suministrarse en conjunto con otro compuesto que estimule el uso de las pobres reservas de ATP del patógeno, como por ejemplo un activador de ATPasas tipo P.

Recientemente, Murphy and Brown han mencionado la importancia de CtpF y CtpG de *M. tuberculosis* como posibles dianas para causar dicho desperdicio de ATP micobacteriano; sin embargo, estas proteínas tienen una considerable similitud con sus ortólogos humanos, lo que exige un gran reto para el desarrollo de inhibidores selectivos para estos transportadores micobacterianos [16].

Para establecer cuál de las ATPasas de *M. tuberculosis* tiene mayor potencial como diana terapéutica, nuestro grupo de investigación está realizando un meta-análisis de múltiples datos de expresión génica y microarreglos previamente publicados. Dichos experimentos han sido realizados en condiciones que simulan la infección tales como hipoxia, inanición, infección en macrófagos y en infección murina. En la actualidad estamos determinando un consenso de las ATPasas tipo P que son más importantes para la supervivencia de la micobacteria durante la infección.

Lo que hemos observado hasta el momento es que varios genes de ATPasas tipo P se sobreexpresan en presencia de sustancias tóxicas como antimicrobianos, S-nitrosoglutatión (GSNO), etanol, H₂O₂ y óxido nítrico [17-19], en bajas tensiones de oxígeno [20, 21], e incluso algunas de ellas se sobreexpresan durante la infección de macrófagos [22].

Por otro lado, para establecer cuál bomba es mejor diana terapéutica es importante tener en cuenta, qué tan divergentes son las ATPasas del bacilo tuberculoso con respecto a las ATPasas tipo P de humanos y de los organismos que hacen parte de microbiota intestinal humana (para evitar toxicidad del medicamento).

Entonces, la idea general del estudio bioinformático es encontrar zonas conservadas en las ATPasas tipo P micobacterianas, que estén ausentes en las ATPasas tipo P de humanos y de organismos simbióticos presentes en la flora intestinal humana. Puntualmente se pretende:

- 1. Buscar similitudes entre Ortólogos cercanos de las ATPasas tipo P de *M. tuberculosis*:** regiones conservadas presentes en ATPasas tipo P tanto de *M. tuberculosis* como de otras Actinobacterias patógenas como por ejemplo: *M. avium*, *M. bovis*, *M. leprae*, *Corynebacterium diphtheriae*, *C. amicolatum*, *C. striatum*, *C. jeikeium*, *C. urealyticum*, *C. xerosis*, *Nocardia asteroides*, *N. brasiliensis* y *Streptomyces somaliensis*. **Set 1.**
- 2. Buscar similitudes entre Ortólogos lejanos de las ATPasas tipo P de *M. tuberculosis*:** regiones conservadas presentes en ATPasas tipo P de humanos, de otros organismos superiores y de organismos presentes en la flora del tracto digestivo humano (*Homo sapiens*, *Mus musculus*, *E. coli* y *E. faecalis*, entre otros. **Set 2.**
- 3. Buscar diferencias entre los sets 1 y 2:** ósea establecimiento de diferencias entre las ATPasas tipo P de humanos y las ATPasas tipo P micobacterianas.

Se espera que los dominios de las ATPasas tipo P micobacterianas identificados como posibles blancos moleculares sean “drogables”, ósea que puedan ser inhibidos específicamente por compuestos dirigidos a ellos. En ello, las ATPasas tipo P tienen una aparente ventaja y es que los sitios de unión a ATP o fosforilación suelen ser buenas dianas, además que en la actualidad existe mucha investigación respecto al desarrollo de inhibidores de quinasas [23].

Los blancos identificados serán sometidos a un análisis posterior usando Docking molecular, para buscar potenciales inhibidores.

Referencias

-
1. Pethe, K., et al., *Isolation of Mycobacterium tuberculosis mutants defective in the arrest of phagosome maturation*. Proc Natl Acad Sci U S A, 2004. **101**(37): p. 13642-7.
 2. Cosma, C.L., D.R. Sherman, and L. Ramakrishnan, *The secret lives of the pathogenic mycobacteria*. Annual Reviews Microbiology, 2003. **57**: p. 641-676.
 3. Sundaramurthy, V. and J. Pieters, *Interactions of pathogenic mycobacteria with host macrophages*. Microbes and Infection, 2007. **9**: p. 1671-1679.
 4. But, P.G., et al., *Intracellular Transformation of Phagosomes*. Biology Bulletin, 2004. **31**(6): p. 564-567.
 5. Kelley, V.A. and J.S. Schorey, *Mycobacterium's arrest of phagosome maturation in macrophages requires Rab5 activity and accessibility to iron*. Mol Biol Cell, 2003. **14**(8): p. 3366-77.
 6. Botella, H., et al., *Mycobacterial p(1)-type ATPases mediate resistance to zinc poisoning in human macrophages*. Cell Host Microbe, 2011. **10**(3): p. 248-59.
 7. Palmgren, M.G. and P. Nissen, *P-type ATPases*. Annu Rev Biophys, 2011. **40**: p. 243-66.
 8. Chan, H., et al., *The p-type ATPase superfamily*. J Mol Microbiol Biotechnol, 2010. **19**(1-2): p. 5-104.
 9. Kuhlbrandt, W., *Biology, structure and mechanism of P-type ATPases*. Nat Rev Mol Cell Biol, 2004. **5**(4): p. 282-95.
 10. Thever, M.D. and M.H. Saier, Jr., *Bioinformatic characterization of p-type ATPases encoded within the fully sequenced genomes of 26 eukaryotes*. J Membr Biol, 2009. **229**(3): p. 115-30.
 11. Axelsen, K.B. and M.G. Palmgren, *Evolution of substrate specificities in the P-type ATPase superfamily*. J Mol Evol, 1998. **46**(1): p. 84-101.
 12. Novoa-Aponte, L., et al., *In silico Identification and characterization of the ion transport specificity for P-type ATPases in the Mycobacterium tuberculosis complex*. BMC Struct Biol, 2012. **12**(1): p. 25.
 13. Rathnayake, I.V.N.M., M.; Bolan, N. and Naidu, R. , *Tolerance of Heavy Metals by Gram Positive Soil Bacteria*. World Academy of Science, Engineering and Technology, 2009. **53**: p. 1185-1189.
 14. Ward, S.K., et al., *CtpV: a putative copper exporter required for full virulence of Mycobacterium tuberculosis*. Mol Microbiol, 2010. **77**(5): p. 1096-110.
 15. Andries, K., et al., *A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis*. Science, 2005. **307**(5707): p. 223-7.
 16. Murphy, D.J. and J.R. Brown, *Identification of gene targets against dormant phase Mycobacterium tuberculosis infections*. BMC Infect Dis, 2007. **7**: p. 84.
 17. Waddell, S.J., et al., *The use of microarray analysis to determine the gene expression profiles of Mycobacterium tuberculosis in response to anti-bacterial compounds*. Tuberculosis (Edinb), 2004. **84**(3-4): p. 263-74.
 18. Kendall, S.L., et al., *The Mycobacterium tuberculosis dosRS two-component system is induced by multiple stresses*. Tuberculosis (Edinb), 2004. **84**(3-4): p. 247-55.
 19. Hampshire, T., et al., *Stationary phase gene expression of Mycobacterium tuberculosis following a progressive nutrient depletion: a model for persistent organisms?* Tuberculosis (Edinb), 2004. **84**(3-4): p. 228-38.
 20. Bacon, J., et al., *The influence of reduced oxygen availability on pathogenicity and gene expression in Mycobacterium tuberculosis*. Tuberculosis (Edinb), 2004. **84**(3-4): p. 205-17.
 21. Sherman, D.R., et al., *Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha -crystallin*. Proc Natl Acad Sci U S A, 2001. **98**(13): p. 7534-9.
 22. Schnappinger, D., et al., *Transcriptional Adaptation of Mycobacterium tuberculosis within Macrophages: Insights into the Phagosomal Environment*. J Exp Med, 2003. **198**(5): p. 693-704.
 23. Fabian, M.A., et al., *A small molecule-kinase interaction map for clinical kinase inhibitors*. Nat Biotechnol, 2005. **23**(3): p. 329-36.

Cooperación Internacional

Miembro fundador del PanAmerican Bioinformatics Institute

<http://panambioinfo.org/>

La misión del Instituto Panamericano de Bioinformática es la de facilitar la salud pública y el desarrollo económico en las Américas a través de la bioinformática educación, investigación y extensión.

Ver publicaciones en cooperación con investigadores en Colombia.

PanAmerican Bioinformatics Institute

Leveraging science for health and development across the Americas

The mission of the **PanAmerican Bioinformatics Institute (PABI)** is to facilitate **public health** and **economic development** in the Americas through **bioinformatics education, research** and **outreach**. The **PABI** was originally formed as a collaborative effort between faculty from Universidad Nacional de Colombia and the Georgia Institute of Technology in the USA; we are currently made up of a growing network of faculty and student researchers in Colombia and the United States. In order to achieve our mission, we are engaged in a series of activities centered on specific bioinformatics and genomics research themes that are directly related to public health and economic development in Latin America. Members of the **PABI**: 1) offer bioinformatics workshops and short courses for students and faculty in Colombia, 2) facilitate student exchanges between host institutions and laboratories in Colombia and the USA, and 3) engage in collaborative research efforts in bioinformatics and genomics that bring together laboratories from Colombia and the USA. The goal of all of these efforts is to help create the **local human capacity** needed to deal with **region-specific challenges** for **public health** and **economic development** across the Americas. In order to initially focus the work of the **PABI**, we have chosen to build on standing collaborative relationships between Colombia and the USA. This focus takes advantage of strong existing ties between laboratories from the two nations and also benefits from the distinguishing characteristics of the Colombian people, a few of which are enumerated below. It is anticipated that this two country model will ultimately be extended across the Americas.



Why Bioinformatics?

The biological sciences are undergoing a revolution in experimental technology that is resulting in massive data sets of unprecedented breadth and depth. For instance, it is now possible to characterize the sequence of an entire human genome in a single day for only a few thousand dollars. This revolution and the resulting deluge of data present profound challenges as well as unique opportunities.

The bottleneck in biological research occurs more often than ever at the level of computational analysis of large-scale data sets. Accordingly, bioinformatics education and research will only become more important as computational work becomes an indispensable part of the life sciences. Aside from the importance of the fields, bioinformatics and computational biology are marked by features that make them uniquely accessible research endeavors. First and foremost, bioinformatics and computational biology research can be relatively inexpensive compared to experimental biology. Substantial contributions to research in computational biology, and even fundamental discoveries, can be made using only a desktop computer and an internet connection. Even high-performance computing can be done on the cheap now thanks to the advent of cloud computing. In addition, bioinformatics has benefited tremendously from the open source software movement. Virtually all major state-of-the-art bioinformatics software applications are made freely and readily available to academic researchers. Thus, as opposed to expensive facilities, equipment and reagents, the major limiting factors for bioinformatics and computational biology are the enthusiasm, skills and imagination of the researchers engaged in the work. Fortunately, these human characteristics can be found in abundant supply among the people of Colombia. In short, because of its relatively low cost and the emphasis on human capital, we are convinced that the field of **bioinformatics can be a paradigm of science as an engine for human development** in the Americas.

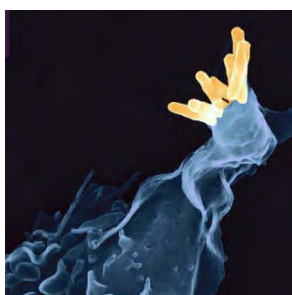
Why Colombia?

The South American nation of Colombia has undergone a transformation over the last decade and is now poised to emerge as a major player in the hemisphere. The country that was plagued by years of internal armed conflict has seen marked increases in security and stability along with continued economic growth. BusinessWeek recently declared Colombia as "the most extreme emerging market on Earth" and the New York Times named the country as one of its top travel destinations for 2010. Nevertheless, the reputation of Colombia has not yet caught up with the reality of increased stability and productivity that exist on the ground. As such, there are abundant opportunities for outreach and international collaboration with Colombia that have yet to be tapped. Colombia also possesses a world-class system of universities and a highly educated populace. PABI faculty know from experience that Colombian students are passionate about their education and relish the opportunity to engage with students and faculty from abroad. Indeed, human capital represents the richest resource in Colombia and throughout all of Latin America for that matter. The recent emergence of a more stable Colombia, together with the longer standing tradition of academic excellence, make the timing ideal for investment in bioinformatics education and research in Colombia.



Specific Challenges

The **PanAmerican Bioinformatics Institute** aims to use **bioinformatics education, research and outreach** to help create the **local human capacity** needed to deal with **region-specific challenges** for **public health and economic development** in Colombia. To address these broad goals in a targeted fashion, we have chosen to initially concentrate the efforts of PABI on two specific research themes: 1) computational genomics of *Mycobacterium colombiense* and 2) transcriptomics of *Physalis peruviana* L.



Mycobacterium colombiense is a novel species of the Mycobacterium genus recently identified in Colombian HIV/AIDS patients. Mycobacteria are a leading cause of death from bacterial infections worldwide and a major challenge for public health in developing countries. The public health significance of this particular bacterial pathogen is further underscored by the fact that all Colombian patients who presented *M. colombiense* caused mycobacteremia died from the infections. PABI affiliated investigators are spearheading a genome sequencing and analysis project for *M. colombiense*.

Patient samples and genomic DNA will be isolated by members of the Martha Murcia laboratory at the Universidad Nacional de Colombia, genome sequencing will be done by the group of Timothy Read at Emory University and genome analysis will be performed as a distributed collaborative effort among PABI member laboratories. Students will be trained in the use of the latest computational genomics tools and techniques using real-world sequence data generated as part of this project.



***Physalis peruviana* L.** is a member of the plant family Solanacea and closely related to tomato and potato. *P. peruviana* L. is indigenous to the high Andes region in South America, and in Colombia it is called by the indigenous name Uchuva. Long known for its medicinal properties, presumably based on demonstrated anti-inflammatory and anti-oxidant properties, Uchuva has recently become an important economic crop in Colombia. PABI member investigators, led by Dr. Leonardo Mariño-Ramírez, are characterizing and analyzing the transcriptome of several varieties of Uchuva. One of the goals of this work is to identify molecular markers associated with disease resistance.

The analysis of the transcriptome data, along with the generation of a *P. peruviana* L. database that will host and distribute sequence data and associated annotations, will be performed by PABI affiliated students under the guidance of PABI faculty in Colombia and the US. A bioinformatics workshop where *P. peruviana* L. transcriptome sequence data will be actively analyzed will be held at Georgia Tech in 2010.



Leonardo Mariño-Ramírez
 Profesor ad-honorem
 Universidad Nacional de Colombia
<http://www.ncbi.nlm.nih.gov/CBBResearch/Marino>



King Jordan
 Associate Professor
 School of Biology
<http://jordan.biology.gatech.edu>

PABI Founders

The **PanAmerican Bioinformatics Institute** was founded by Dr. Leonardo Mariño-Ramírez from Universidad Nacional de Colombia and Dr. King Jordan from the Georgia Institute of Technology in the USA. Drs. Mariño-Ramírez and Jordan have been close research collaborators since 2002 and have worked together on bioinformatics education and outreach in Colombia since 2005.

PABI Activities

The **PanAmerican Bioinformatics Institute's** public health and economic development missions are addressed through an ongoing series of:

1. bioinformatics workshops and short courses for students and faculty in Colombia
2. student exchanges between host institutions and laboratories in Colombia and the USA
3. collaborative research efforts in bioinformatics and genomics

Past activities: PABI faculty have taught in four bioinformatics summer workshops or short courses in Colombia since 2005. In the last year, four Colombian graduate students have interned in US host laboratories

Ongoing activities: PABI faculty and students are involved in targeted collaborative research projects on the genomics of *Mycobacterium colombiense* and the transcriptome analysis of *Physalis peruviana* L.

Upcoming activities: There will be a bioinformatics workshop involving **PABI** faculty along with Colombian and US students at Georgia Tech in December 2010. There will be a applied computational genomics summer Course in Colombia in July 2011.

Giving Opportunities

Several gift opportunities are available to support the **PanAmerican Bioinformatics Institute**. Our key funding needs are outlined below (shown in US dollars):

SUPPORTING INTERNATIONAL WORKSHOPS

Bioinformatics workshop , Georgia Tech, December 2010	\$15,000
Applied computational genomics course, Colombia, July 2010	\$25,000

Funds will support the continuation of workshops and short courses that bring together faculty and students from Colombia and the USA

INTERNATIONAL GRADUATE FELLOWSHIPS

Colombian graduate student exchange program	\$40,000
US graduate student instructional fellowship	\$20,000

Colombian students will study in US host laboratories and US students will travel to Colombia to engage in instructional activities and research

SUPPORTING COLLABORATIVE RESEARCH EFFORTS

<i>Mycobacterium colombiense</i> genome project	\$50,000
---	----------

Funds will support the genome sequencing and analysis of this pathogenic bacterium isolated from Colombian HIV/AIDS patients